IMPROVING THE QUALITY OF SPEECH IN NOISY ENVIRONMENTS

A Thesis
Presented to
The Academic Faculty

by

Devangi N. Parikh

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December 2012

IMPROVING THE QUALITY OF SPEECH IN NOISY ENVIRONMENTS

Approved by:

Professor David V. Anderson, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Pamela T. Bhatti
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Mark A. Clements
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Branislav Vidakovic
School of Biomedical Engineering
Georgia Institute of Technology

Professor James H. McClellan
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: 31 August 2012

To my parents, Ragini and Nikunj,

and my sister, Dulari.

ACKNOWLEDGEMENTS

As I approach the end of my tenure at Georgia Tech, I take this opportunity to thank everyone who made this achievement possible.

Foremost, I must thank my advisor Dr. Anderson, who decided to hire me as a research assistant 6 years ago. That is when my journey began. His guidance and constant support through out my research helped me sail through this experience easily.

Secondly, I would like to thank my family—my father, Nikunj, for teaching me that hard work is always rewarded and to never give up; my mother, Ragini, for being the pillar of support I needed; my sister, Dulari, for motivating me to strive for the best.

At my stay at Georgia Tech, I made a lot of friends—the friends that helped maintain my sanity, the friends that were there when I needed to vent, the friends that were around to answer my silly technical questions, and the friends that taught me some important life lessons. They have all truly enriched my experience at Georgia Tech.

I must thank my reading committee Dr. Clements, and Dr. McClellan. Their input throughout my career at Georgia Tech encouraged me to strive for excellence as a researcher. Moreover, I'd like to thank Dr. Bhatti and Dr. Vidakovic for their valuable input and taking time out of their busy schedules to serve on my committee.

Lastly, I am grateful for Fr. Morondo, my high school Biology teacher, for being my role model during the early years of my academic career. From him I learned that knowledge is far beyond just text book theory. He is the inspiration of my passion for teaching.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

To suppress the background noise present in a noisy speech signal, typically, the noisy signal is filtered with an adaptive filter that varies with time. In most noise-suppression algorithms, the noisy signal is transformed into the frequency domain using a fast Fourier transform. For a particular time frame of the noisy signal, the adaptive filter suppresses the frequencies of the signal that contains noise. This filter adapts over time such that the speech, which varies with time is not altered, and the noise is attenuated. This adaptive filter can also be considered as a frequency-selective time-varying gain that is applied to the noisy signal to suppress noise. The adaptive filter, in other words the noise-suppression gain, is then obtained by optimally estimating the frequency spectrum of the clean speech.

When a noise-suppression gain is applied to the signal, it modulate the signal and this modulation creates distortion product terms. The frequency and energy of these product terms depend on the rate of change and the peak-to-peak amplitude of the gain G, respectively. The frequency and energy of the product terms determine if they are perceivable as artifacts and distortion in the final processed speech. Since, most traditional noise-suppression algorithms are not developed keeping these constraints in mind the processed output typically contains audible artifacts. Efforts have been made to reduce this artifacts. While these methods work well, they become computationally complex and require significant tweaking.

In this thesis, we are interested in processing signals that are meant to be heard by humans, and hence we approach the noise-suppression problem from a perceptual perspective. We develop a noise-suppression paradigm that is based on a model of the human auditory system, where we process signals in a way that is natural to the human ear. Under this paradigm, we transform an audio signal in to a perceptual domain, and processes the signal in this perceptual domain. This approach allows us to reduce the background noise and the audible artifacts that are seen in traditional noise-suppression algorithms, while preserving

the quality of the processed speech. We develop a single- and dual-microphone algorithm based on this perceptual paradigm, and conduct subjecting tests to show that this approach outperforms traditional noise-suppression techniques. Moreover, we investigate the cause of audible artifacts that are generated as a result of suppressing the noise in noisy signals, and introduce constraints on the noise-suppression gain such that these artifacts are reduced.

CHAPTER I

INTRODUCTION

Telecommunication devices have become an integral part of our lives, and the effectiveness of these devices to carry out conversations with ease depends on how much background noise can be suppressed without altering the quality of speech. Noise-suppressed speech is especially important with audio codecs, since the codec may not be able to code the speech correctly if the speech is buried in noise or is distorted. Moreover, speech signals that have been preprocessed with noise-suppression algorithms can improve the performance of speech recognition algorithms. While speech enhancement and noise suppression has been a topic of research for the past several decades, the problem of background noise and processed speech that sounds unnatural still exists.

Traditionally, the noise-suppression problem is approached from a mathematical optimal-estimation aspect. The speech signal is typically analyzed into its temporal-spectral components using a short-time Fourier transform. A noise-suppression gain is obtained using an optimal estimator, and assumes a statistical model to describe the spectral coefficients of the signal. The power spectrum of the clean speech is then estimated using the assumed statistical model. Based on the estimated power spectrum, the noise-suppression gain is calculated for each frame by minimizing the noise energy or maximizing the signal-to-noise ratio (SNR). While the output may be the mathematically optimal solution, the speech processed with such a gain, may not *sound* natural to the human ear, that is, the speech may be distorted and may also contain audible artifacts. One of resulting artifacts of this type of processing is called musical noise. Musical noise is heard as random tones that appear at different times. This artifact is generated by the gain suppressing parts of the spectrum by varying amounts over time.

In order to solve the noise-suppression problem assumptions are made regarding the noisy speech signals so that the clean speech spectrum can be optimally estimated. The

speech signal is typically assumed to be stationary over a segment of time, and the spectral coefficients are typically assumed to be Gaussian. In reality, these assumptions are not true, but are made to simplify the noise-suppression problem. In most cases, an estimate of the noise-suppressed speech is obtained by minimizing the noise energy present in the noisy speech and not by optimizing for the perceptual quality of speech. Hence, the output may not necessarily sound natural to the human ear.

To address these problems, often, the optimal gain is abandoned. The gain is smoothed over time and frequency to make sure the gain is not changing rapidly. This gain smoothing reduces the audibility of most of the artifacts. Moreover, certain principles of the human auditory model, such as the principle of masking, and the threshold of hearing, may also be incorporated in the noise-suppression gain. The noise-suppression gain is modified such that the audible artifacts are either masked by other louder sounds or are lower than the threshold of hearing. While the speech that is processed using these approaches sounds reasonable to the human ear, the noise suppression algorithm is often very complex, and requires substantial tweaking.

Early analysis of the auditory system showed that the Fourier transform is an appropriate domain to analyze sound signals. However, further research showed that the uniform frequency analysis using an FFT may not be the best representation to understand how the auditory system works. Nobel prize winner for his discoveries of the physical mechanisms of the cochlea, von Békésy, in a posthumously published article in 1974 remarked, "*In time, I came to the conclusion that the dehydrated cats and the application of Fourier analysis to hearing problems became more and more a handicap for research in hearing*" [**?**]. Hence, processing signals based on Fourier analysis may not be the most appropriate approach if the processed signals are meant to be heard by the human ear.

In this thesis, for speech-processing algorithms especially noise-suppression techniques, we abandon the traditional Fourier analysis and use a transform that is more closely related to the auditory system to analyze the signal. Moreover, we develop a noise-suppression rule that is in conjunction with the human auditory perceptual system. This change in speech-processing paradigm will help us to develop algorithms in which the processed output will

have a better perceptual quality than traditional methods. In the next section, we will briefly discuss the previous work that motivates the research in this thesis.

1.1 Principles of Speech Perceptual Processing

As signal processing researchers, there are two approaches of auditory perceptual processing—mimicking the auditory system, and processing signals that are meant to be perceived by the auditory system. In the first approach—mimicking the auditory system, we are interested in micro-modeling the auditory system so that we can understand how signals are analyzed and perceived by the human ear. However, in this approach, just like the auditory system, the signals are analyzed and not synthesized. Hence, such micro modeling may not be invertible. While, in the second approach, processing signals that are meant to be perceived by the auditory system, we are interested in conceptually understanding how the auditory system perceives signals. This understanding is then applied to process signals that are meant to be perceived by the auditory system. In this approach, the auditory system is macro modeled so that the analysis can be inverted to reconstruct the signal. In Chapter 2, we will see how these aspects of perceptual processing are used in speech-processing applications.

In this thesis, we are interested in suppressing noise present in noisy speech samples, which is meant to be heard by humans. We are focused on improving the perceptual quality of the speech and are not interested in improving quantitative metrics of signal-to-noise ratio (SNR), coding efficiency, error in the spectral magnitudes, and speech-recognition rates. In particular, we are interested in processing the signal such that the speech signal remains untouched while the background noise sounds less loud.

Christiansen showed that the quality of the processed speech can be improved over traditional FFT-based methods if the signal is processed in the perceptual or loudness domain [14]. The signal can be transformed into the loudness domain by analyzing the signal using a digital hearing model. The digital hearing model used in this type of processing must have an inverse so that the signal can be synthesized after the processing.

Ravindran proposed speech features that are inspired by the human auditory perceptual model [48]. He showed that these features performs better than traditional methods, especially if background noise is present. Moreover, Ravindran showed that by expanding the dynamic range of the signal envelope that is extracted from the subbands, the amount of noise present in the signal and hence in the speech features can be reduced. This dynamic range expansion improves the performance of the speech recognition system in noisy conditions.

When audio signals are processed, a time-varying gain is applied to the signal. This time-varying gain modulates the signal and product terms are generated. These product terms are usually audible to the human ear as artifacts. However, Anderson explained that if the signals are processed in the auditory critical bands such that the product terms remain within the critical band, then the artifacts will not be audible [5].

Based on the principle findings of these works, we develop a noise-suppression algorithm that transforms an audio signal in to the perceptual domain. This transformation is obtained by decomposing the signals in to the auditory critical bands and approximately mimicking the effect of the outer hair cells during audition. Such processing allows us to suppress the background noise without altering the quality of speech.

In any noise suppression algorithm determining when speech is present and noise is present is a major challenge. From a signal processing perspective, the shape and energy of the envelope can be used to determine if speech is present at a particular time. However, in humans the ability of our auditory system to localize sound sources helps us to distinguish what sounds are of interest to us and what sounds are noise. The noise sources are automatically filtered by the auditory system. The human auditory systems uses cues such as the time- and level-differences of the sound reaching both ears, and spectral information to localize sources present in space. While duplicating the operations of the auditory system and the brain that localize sounds is beyond the scope of this thesis, we can use techniques that mimic the task of separation of spatially disparate sources to estimate the noise present in the noisy speech. In this thesis, we use a blind source separation (BSS) to separate the sources. We combine such source separation techniques with the perceptually-motivated

noise-suppression techniques to further improve the noise-suppression capabilities of our algorithms.

We also show that we can combine perceptual-model-based processing with mathematical optimization techniques to obtain a noise-suppression rule. Such a combination formalizes the approach of calculating the parameters of the noise suppression gain.

1.2 Thesis Organization

This thesis is organized as follows:

Chapter 2 In this chapter, we present an overview of single-microphone noise-suppression techniques. We briefly discuss blind source separation algorithms, which separates sources that are in spatially disparate locations from a given mixture. Moreover, the human auditory perceptual model, and how these models are incorporated in speech-processing techniques is also described in this chapter.

Chapter 3 Here, we present a single-microphone noise-suppression system that is based on the human auditory perceptual model. The signal is decomposed into critical bands from which an envelope is extracted. The noise-suppression gain is calculated on the basis of this envelope. The gain is such that the background noise in the speech is suppressed by expanding the dynamic range of each critical band signal. This proposed system will provide a basic frame work of signal analysis and synthesis that will be used in the following chapters.

Chapter 4 In this chapter, we modify the perceptual speech enhancement technique described in Chapter 3, so that it can be used a post-processing technique for blind source separation. Moreover, in this chapter, we discuss the experimental setup used for testing our algorithm.

Chapter 5 In this chapter, we modify the perceptual post-processing technique so that it can be implemented in real time. We will see that this real-time implementation generates audible musical noise in the processed signal. We, then, describe a technique

used to reduce the artifacts that are generated by updating the parameters of the gain on a frame-by-frame basis.

Chapter 6 In this chapter, we derive a noise-suppression rule by combining the frame work of the perceptual signal analysis and mathematical optimal estimation techniques.

Chapter 7 Here, we investigate the cause of musical noise and suggest constraints that can be applied to the proposed noise-suppression algorithm to reduce the amount of perceived musical noise.

Chapter 8 Concluding remarks are made in this chapter. We present a summary of contributions of this research as well as future work based on this thesis.

CHAPTER II

BACKGROUND

In this chapter, we discuss the basics of speech-enhancement algorithms, the human auditory system, and models that approximate the auditory system.

2.1 Speech Enhancement and Blind Source Separation

In this section, we explain the basic concepts of speech enhancement, and blind source separation.

2.1.1 Speech Enhancement

A general block diagram of a noise-suppression algorithm is shown in Figure 1. The noisy signal is analyzed and decomposed into the frequency domain. This frequency decomposition is usually done using the fast Fourier transform (FFT). The amount of noise present in each band is then estimated. From this noise estimate, a gain G is calculated such that the noise present in the signal is suppressed.

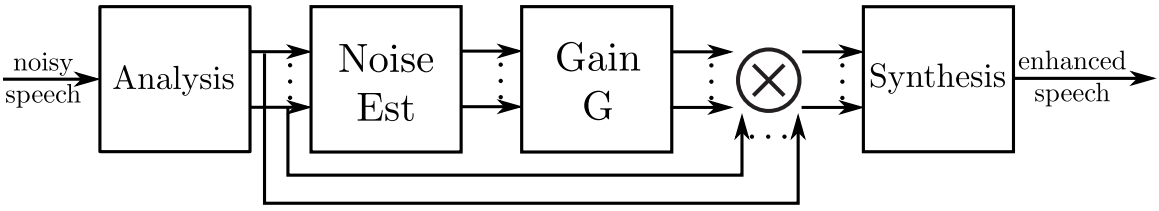


Figure 1: General block diagram of a typical noise suppression algorithm.

Different algorithms calculate the gains based on different concepts. Spectral subtraction algorithms, such as the one described in [10], assume that the noise is additive. The noise-suppressed speech spectrum is obtained by estimating the spectrum of the noise present in the signal and subtracting it from the noisy-speech spectrum. The drawback of this method is that we do not have prior knowledge about the noise and hence we cannot estimate it correctly. The task of estimating the noise spectrum in non-stationary noise is more difficult

since the noise is constantly changing. In [27], a better estimate of the noise is obtained by estimating the noise spectrum independently in each band.

The spectral-subtraction algorithms often leave spurious peaks in the spectrum because of an incorrect estimate of the noise. These peaks, present in different locations in the noise-suppressed speech, result in artifacts called musical noise. Modifications can be made to the spectral-subtraction algorithm to reduce the resulting musical noise. Spectral over subtraction, in which the noise spectrum is over estimated, is one of the techniques used to reduce the musical noise in the output [9].

Wiener filters are widely used for noise suppression [32]. The noise-suppression filter coefficients are obtained by minimizing the mean squared error between the estimate of clean speech and original clean speech. These filter coefficients are a function of the estimates of power spectral density of the clean speech and the noise. The clean-speech spectrum can be estimated iteratively since the clean-speech spectrum is not available. Non-linear estimators can be used to estimate the magnitude spectrum of the signal from the probability density function (PDF) of the discrete Fourier transform (DFT) coefficients of both the noise and speech.

The Ephraim-Malah suppression rule is minimum mean-squared error spectral amplitude estimator [18]. The magnitude spectrum of the clean speech is a non-linear least squares estimate given the noisy signal. The noise suppression is derived assuming that the FFT coefficients are independent Gaussian random variables. This results in a noise suppression rule that is a function of the *a priori* and *a posteriori* SNR estimates. The *a priori* SNR is calculated using a decision-directed approach such that the estimate is smoothed over time. This smoothing ensures the gain does not change rapidly during the noise-only segments, and hence reduces the musical noise artifacts.

2.1.2 Blind Source Separation

Blind source separation (BSS) is a technique that separates the sources from a given set of mixtures. The source separation is "blind" in the sense that the original source and mixing environment are not available. BSS does not recover the original sources exactly, but it

recovers the sources up to a filtered and scaled version of the original signal. If we can find a set of sources that solves the unmixing system, then a permuted set of these sources also solves the system. Hence, the output of the BSS algorithm can be a permuted version of the original sources. Various techniques have been developed to solve the source-separation problem. A detailed survey about these source-separation methods for mixtures obtained in real environments can be found in [34, 43, 54].

Mixtures that contain only scaled version of the inputs are called instantaneous mixtures. The sources from these mixtures can be separated using independent component analysis (ICA) [15]. However, in real acoustical scenarios the mixtures are a filtered sum of the sources. These mixtures are referred to as convolutive mixtures. A linear convolution in the time domain can be represented as a multiplication in the frequency domain. The convolutive mixtures can be simplified to an instantaneous mixture in each frequency bin by transforming the mixtures to the frequency domain. The sources can then be separated using frequency-domain ICA (FDICA). Details of this implementation can be found in [6, 11, 25, 50]. However, since BSS is performed on each frequency bin independently, there is a scaling and permutation ambiguity across frequency bins. Prior knowledge about the sources or the mixing filters can be used to solve this permutation problem.

Source-separation algorithms are developed based on the assumptions made regarding the sources that are mixed. If the sources that are mixed are assumed to be statistically independent of each other, then all the cross-moments between the sources are zero. Therefore, by minimizing all the cross-moments of the mixtures to zero, we can separate the sources. In certain cases, it may not be necessary to minimize all the cross-moments of the mixtures to achieve separation. A cumulant-based method in the frequency domain is given in [29]. Convolutive mixtures can also be separated by minimizing the fourth-order cumulants [11]. The second-order statistics in the frequency domain can be represented by the cross-power spectrum. Separation can be achieved by minimizing the cross-power spectrum between the sources. This method is described in [25].

Statistical independence of the sources can be expressed in terms of the PDF of the signals. If the sources are independent, then the mutual information between these sources

is zero. Thus, the speech sources can be separated by minimizing the mutual information between the sources; this is equivalent to maximizing the entropy [8]. However, the PDF or the cumulative density function (CDF) of the signal must be modeled appropriately. The CDF of speech signals can be approximated by passing the source through a non-linear function such as hyperbolic tangent (tanh). Algorithms based on this method are described in [2, 16].

If the sources that have been mixed are assumed to be sparse in the time-frequency (T-F) domain, then at each T-F point only one source is dominant. This assumption holds true for speech signals. The T-F points can be projected onto a space where all the T-F points associated with each source form a cluster. This projection can be based on the direction of arrival of the sources. From the clustered points, we can generate a mask that retains one source and zeros out the T-F points associated with the other sources. Algorithms based on this idea are described in [7, 31].

2.2 Understanding the Human Auditory System

The human ear is divided into three parts – outer ear, middle ear and inner ear. The outer ear is responsible for channeling the sound to the latter parts of the ear and providing important cues for localization of the sound source. The middle ear, which consists of three bones, provides impedance matching between the sound that is incident on the ear and the inner ear.

The conversion of the mechanical sound-pressure waves to the electrical neural firing occurs in the inner ear. The cochlea within the inner ear is the main organ responsible for this conversion. The basilar membrane and organ of Corti reside in the cochlea. The basilar membrane vibrates with the sound-pressure waves. Different regions of the basilar membrane respond to different frequencies because of the difference in size and stiffness along the length of the basilar membrane. Thus, any given frequency will excite one particular location in the cochlea. The frequency decomposition of the cochlea can be described psychologically by critical bands. Signals that have the same energy level and fall within a critical band are perceived to have the same loudness. Moreover, if a signal and masker are

present simultaneously, then the frequencies of the masker that fall within the same critical band as the signal contribute to the masking of the signal.

The hair cells that reside on the basilar membrane in the organ of Corti regulate the dynamic range of the signal intensity and convert the signal into neural responses. There are two types of hair cells—inner and outer hair cells. The outer hair cells provide a mechanical feedback to the basilar membrane to modify the vibration patterns of the basilar membrane. The feedback is such that low-intensity signals are amplified while for high-intensity signals the amplification is inhibited. This feedback non-linearly compresses the large input dynamic range of signal intensities the ear can hear to a small dynamic range of the allowable basilar membrane vibrations [36].

The inner hair cells transduce the motion of the basilar membrane into neural pulses. These neural signals reach the auditory cortex. The inner hair cells and the auditory nerve spontaneously fire in the absence of sound. For the sound to be perceived by the brain, the sound must cause the nerves to fire at a rate higher than the spontaneous-firing rate. The nerves tend to fire at a particular phase of the input simulating signal.

The non-linear compression of the outer hair cells, the rate of firing of the inner hair cells, and an exponentiation process in the higher auditory centers of the brain describe the perceived loudness. The perceived loudness can be characterized by phon or sone scales. More information about these scales can be obtained from [36] and [58].

Next, we explain how different aspects of the perceptual auditory system are used in speech processing techniques.

2.3 Human Perception in Speech Processing

Algorithms inspired by the human auditory system are seen in different areas of speech processing. For example, perceptual models are used in audio- and speech-coding algorithms [26]. The amount of audio information that has to be coded can be drastically reduced by discarding the sounds that are masked by other sounds in the audio.

To reduce the artifacts and distortions in the noise-suppressed speech a limit is enforced on the gain G so that it does not distort the speech. This constraint limits the amount

of noise suppression that can be achieved but, however, does not necessarily eliminate the artifacts and distortions. Researchers use the masking property of the human auditory system to modify the noise-suppression gain so that the resulting artifacts are masked. Virag [57] proposes a method to calculate the over-subtraction, and noise-floor parameters of the spectral-subtraction algorithm adaptively on the basis of the masking thresholds of the ear. These parameters are set such that an optimal balance between the noise reduction and speech quality is achieved.

Gustaffson *et al.* propose to mask the distortions of the residual noise while allowing variable speech distortions [21]. This is achieved by calculating the spectral-subtraction gain is calculated such that the residual noise is exactly at the masking threshold. The gain set in this way not only reduces the distortion in the noise but also minimizes the speech distortions. In [53], the spectral-subtraction levels are calculated from a psychoacoustic model that evaluates sound quality. Lai *et al.* propose an algorithm that attenuates the noise below the audible threshold [30]. The tonal components of the power spectrum are located and the spectral-subtraction gain is calculated in such a way that the residual noise remains below this masking threshold. Lai *et al.* show that the performance of a standard speech recognizer improves when the speech samples are processed using this psychoacoustic-model-based spectral subtraction. Moreover, masking thresholds are also used to limit the Wiener gain [1].

Psychoacoustics-based processing has also been used for echo cancellation along with noise suppression. A single-channel echo-cancellation system is described in [20], and a multi-channel system is described in [49].

However, most of the noise suppression techniques described above use only one aspect of the perceptual auditory model in conjuction with standard noise suppression to process speech samples. Next, we see how the human auditory system is modeled so that this model can be used for speech enhancement algorithms.

2.4 Modeling the Human Auditory System

The human auditory system can be either micro modeled, where each component of the auditory system is modeled individually, or macro modeled, where the input-output relationship between the sound and perceived quantity is modeled [14]. In the human auditory system, once the speech signals is analyzed by the brain the speech does not have to be resynthesized. Hence, most of the models of the human auditory system are not invertible. However, to use a model of the perceptual auditory system for speech processing, where the output will be perceived by humans, it is important that the model be invertible so that the processed speech can be resynthesized.

Christiansen proposes a macro model of hearing, which maps the physical sound signal to the perceptual domain. This model maps the intensity of the signal to the sone scale. The mapped intensity, which is in the perceptual domain, corresponds to the perceived loudness of the signal [14]. The signal is processed in the perceptual domain and then the signal is synthesized by applying the inverse hearing model. A block diagram of this process is shown in Figure 2.

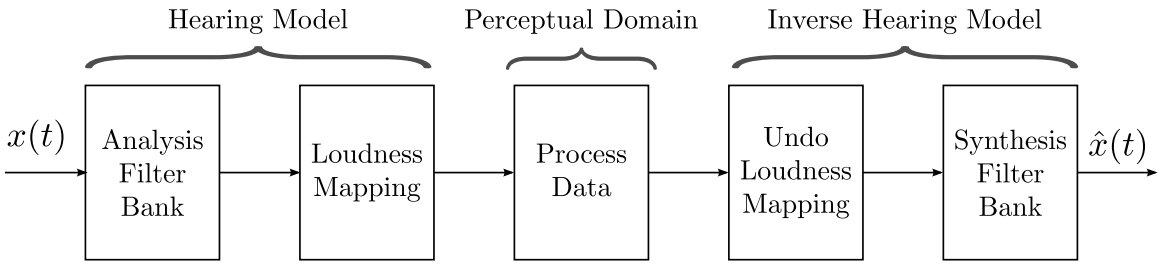


Figure 2: Block diagram of a macro model of the human auditory system. This system was proposed by Christiansen to process speech signals [14].

In [44], Petersen *et al.* also proposes to perform spectral subtraction in the perceptual domain. However, the mapping that is used is based on Steven's power law. This power law maps the input intensity to the perceived loudness (sone scale) by a fixed-exponent power function. It was later found that a conversion from the phon scale to sone scale is more accurate in terms of loudness mapping [3].

Anderson proposes a hearing model that mimics the four main functions of the auditory

system – filtering, dynamic range compression, signal transduction and loudness perception [3]. The simplified model is shown in Figure 3.

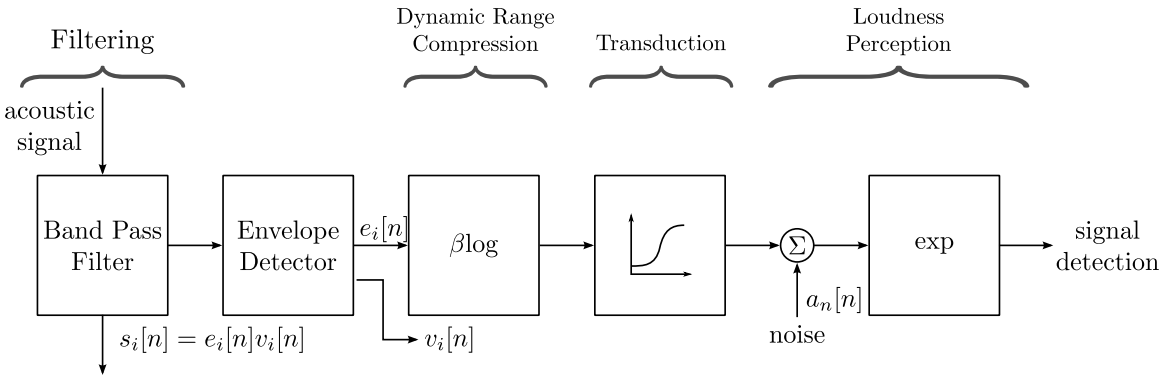


Figure 3: Simplified block diagram of the hearing model for a single channel.

In [3], Anderson also proposes a model for hearing aids. The model is formed by concatenating a normal hearing model with the inverse of the impaired-ear model. The inverse of the impaired-ear model compensates for the impairment of the ear. The sound, thus processed by the hearing-aid model, is perceived in the same way a normal-hearing ear perceives the sound. The cascade of the hearing model followed by an inverse impaired-hearing model reduces to a power function relationship between the input and output sound intensity. This model is show in Figure 4.

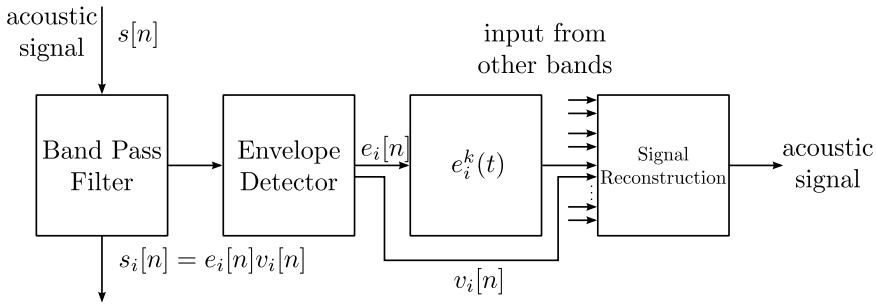


Figure 4: Simplified block diagram of the hearing-aid model for a single channel proposed by Anderson [3].

In this research, we will show that we can obtain noise suppression in noisy signals by using the hearing-aid model shown in Figure 4. Noise suppression can be obtained by adjusting the power function such that the dynamic range is expanded instead of compressed. We will also show that by using two microphones we can improve the amount of noise

suppression obtained by this method without altering the quality of speech. Moreover, we propose that if we replace the power-function block in Figure 4 with other standard noise suppression algorithms we can obtain noise-suppression without altering the quality of speech.

CHAPTER III

PERCEPTUAL NOISE SUPPRESSION

Most of the single-microphone noise-suppression techniques process the noisy signals in the frequency domain using an FFT. A noise-suppression gain is obtained by optimally estimating the clean speech frequency spectrum given the noisy speech frequency spectrum. The noise-suppression gain, when applied to the noisy speech, reduces the background noise. However, the resulting processed speech may contain audible artifacts and there is no guarantee the resulting processed speech will sound natural to the human auditory system. The noise-suppression gain, typically, is tuned such that it is smoothed over time and frequency in an attempt to reduce the audible artifacts and improve the quality of speech. In this chapter, we show that these disadvantages of traditional noise-suppression techniques can be alleviated by developing a perceptual noise-suppression algorithm that mimics the human auditory system. The noise-suppression gain of such a system may require minimal tuning, and the resulting processed speech sounds more natural than the traditional methods.

3.1 Aim of Perceptual Signal Processing

When a signal $x(t)$ is presented to the human auditory system, the auditory system analyzes the signal into critical bands $c_i(t)$. It has been shown that in each critical band the envelope that contains the temporal modulations contains the perceptually relevant information [19]. The brain *hears* a function of $x(t)$ for a given parameter set γ. This function that is perceived by the brain can be modeled as

$$f(x(t), \gamma) = \sum_i \left(\gamma_{i,0} \log_{\gamma_{i,1}} \left(e_i(t) \right) + \gamma_{i,2} \right), \tag{1}$$

where $e_i(t)$ is the envelope of the i-th critical band of $x(t)$, and $\gamma = \{\gamma_{i,j}\}$ is the set of parameters that maps the envelope $e_i(t)$ in the i-th critical band to the corresponding

loudness perceived by the brain. Hence, we can model an acoustic signal $x(t)$ as

$$\begin{aligned} x(t) &= \sum_i c_i(t), \\ &= \sum_i e_i(t) v_i(t), \end{aligned} \tag{2}$$

where $v_i(t)$ is a rapidly varying excitation, and $e_i(t)$ is the slowly varying envelope in the i-th critical band $c_i(t)$ [13].

The model in equation (1) represents the mapping of the signal to the loudness as perceived by the brain. However, equation (1) is just a quantitative approximation of the loudness perception. In reality, the power law may be a more realistic mapping, but is of the same form of (1). The logarithmic approximation is preferred as it results in equations that are computationally simple.

The ear's sensitivity to temporal modulations shows a low-pass characteristics [46, 56]. Viemeister, in [56], showed that the cut-off of this low-pass characteristics is about 100 Hz, while Plomp suggested the cut-off is about 25 Hz [46]. For speech, Houtgast *et al.* measured the modulation index (RMS within 1/3-rd octave bands) for different modulating frequencies and showed that the modulation index for speech signals is the maximum at about 3 Hz [24]. Drullman *et al.* showed that the temporal modulations can be smoothed to 16 Hz without substantially reducing the speech intelligibility [17]. While in [13], for loudness mapping, the authors used an envelope with bandwidth of $\frac{1}{8}$-th of the critical bandwidth. However, Ghitza showed that the carrier in each critical band still contains the modulation information even when the envelope is temporally smoothed to 16 Hz [19]. He also showed that minimum bandwidth of the temporal envelope that contained perceptual information is about one-half of the critical bandwidth.

We are interested in processing the signal $x(t)$ such that the perceptual information carrying $e_i(t)$ is modified and not the rapidly varying carrier $v_i(t)$. This processing can be written as

$$\hat{x}(t) = \sum_i \hat{e}_i(t) v_i(t), \tag{3}$$

where $\hat{x}(t)$ is the processed signal and $\hat{e}_i(t)$ is the modified envelope. By modifying the envelope $\hat{e}_i(t) = g(e_i(t))$, we can process only part of the signal that is perceptually relevant. The envelope modification rule $g(\cdot)$ can be selected such that the processed signal $\hat{x}(t)$ sounds natural to the human auditory system. In this thesis, we select an envelope-modification rule that mimics the non-linear processing of the cochlea. The processed envelope $\hat{e}_i(t)$ of the i-th critical band, can be written as

$$\hat{e}_i(t) = \beta_i (e_i(t))^{\alpha_i}, \tag{4}$$

where β_i and α_i are the parameters of the envelope modification function. This envelope modification follows the multiplicative-AGC model suggested in [13]. In [13], the parameters β_i and α_i are computed such that the dynamic range of the signal is compressed to compensate for hearing loss.

The processed signal $\hat{x}(t)$ can be written as

$$\begin{aligned} \hat{x}(t) &= \sum_i \beta_i (e_i(t))^{\alpha_i} v_i(t), \\ &= \sum_i \beta_i (e_i(t))^{\alpha_i - 1} c_i(t). \end{aligned} \tag{5}$$

According to equation (1), the brain will perceive $\hat{x}(t)$ as

$$f(\hat{x}(t), \hat{\gamma}) = \sum_i \left(\hat{\gamma}_{i,0} \log_{\hat{\gamma}_{i,1}} \hat{e}_i(t) + \hat{\gamma}_{i,2} \right). \tag{6}$$

Substituting equation (4) in equation (6), we get

$$\begin{aligned} f(\hat{x}(t), \hat{\gamma}) &= \sum_i \left(\hat{\gamma}_{i,0} \log_{\hat{\gamma}_{i,1}} (\beta_i (e_i(t))^{\alpha_i}) + \hat{\gamma}_{i,2} \right) \\ &= \sum_i \left(\hat{\gamma}_{i,0} \left(\log_{\hat{\gamma}_{i,1}} (\beta_i) + \alpha_i \log_{\hat{\gamma}_{i,1}} (e_i(t)) \right) + \hat{\gamma}_{i,2} \right) \\ &= \sum_i \left(\hat{\gamma}_{i,0} \alpha_i \log_{\hat{\gamma}_{i,1}} (e_i(t)) + \hat{\gamma}_{i,0} \log_{\hat{\gamma}_{i,1}} (\beta_i) + \hat{\gamma}_{i,2} \right) \\ &= \sum_i \left(\gamma'_{i,0} \log_{\gamma'_{i,1}} (e_i(t)) + \gamma'_{i,2} \right) \\ &= f(x(t), \gamma'). \end{aligned} \tag{7}$$

Hence, when processed signal $\hat{x}(t)$ is presented to the human auditory system, the brain perceives the original unprocessed signal $x(t)$ that is mapped to the loudness domain with another set of parameters γ'. Figure 5 summarizes the aim of perceptual processing.

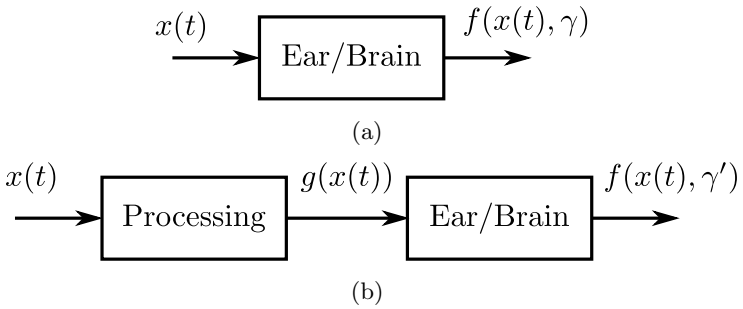


Figure 5: (a) The human ear/brain perceives a function of signal $x(t)$, given a parameter set γ. (b) We are interested in processing the signal $x(t)$ such that the ear/brain perceives the processed signal $g(x(t))$ as a function of the original signal $x(t)$ given a parameter set γ'

3.2 Signal Analysis

We are interested in the implementing the signal model described in equation (2) in a real-time system, hence we switch to a discrete time representation. Equation (2) is written here again in discrete time for a clean speech signal $s[n]$

$$\begin{aligned} s[n] &= \sum_i c_i[n], \\ &= \sum_i e_i[n] v_i[n], \end{aligned} \tag{8}$$

where $v_i[n]$ is a rapidly varying speech excitation, and $e_i[n]$ is the slowly varying speech envelope in the i-th critical band $c_i[n]$.

The critical bands can be obtained by filtering the speech signal with a band-pass filter,

$$c_i[n] = H_i[n] * s[n], \tag{9}$$

where, $H_i[n]$ is the i-th band-pass filter of a constant-Q filter bank. The filter bank is designed to mimic the frequency selectivity of the cochlea. The center frequencies and cut-off frequencies of the band-pass filters [59] are plotted in Figure 6. Each BPF filter is a 4-th order Butterworth filter. The first and last filter of the constant-Q filter bank is a low-pass and high-pass filter respectively. The FIR coefficients of the filter are scaled by 0.75, and the sign of the coefficients are flipped for the odd critical bands. This scaling ensures the summed output response of the filter bank is fairly flat.

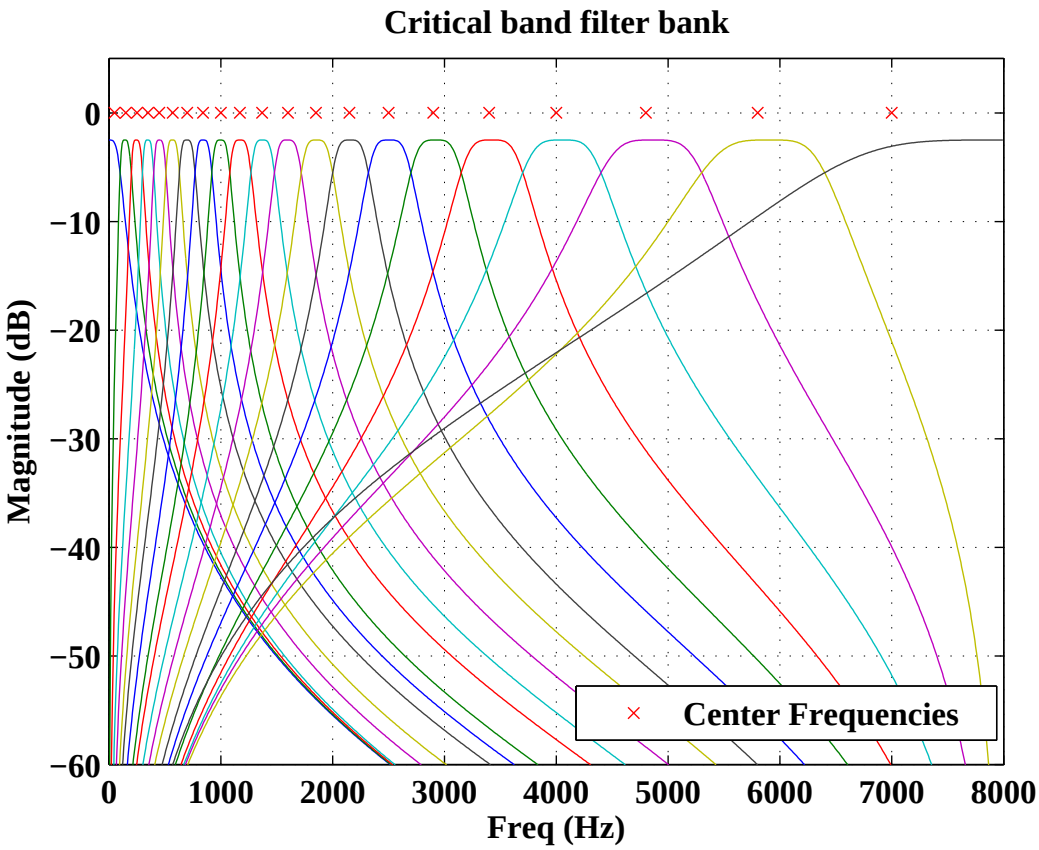


Figure 6: Frequency response of the critical filter bank used in the signal analysis.

In this thesis, the envelope $e_i[n]$ is extracted by full-wave rectifying or squaring and low-pass filtering the i-th critical band $c_i[n]$.

$$e_i[n] = L_i[n] * |c_i[n]|, \tag{10}$$

where, $L_i[n]$ is the low-pass filter (LPF) used to extract the envelope of the i-th critical band. In this thesis, we use a single-pole low pass filter to extract the envelope. The roll-off of a single-pole filter is sufficient to reject the higher frequency content of the critical band. The cut-off frequency of the LPF is set to a fraction of the bandwidth of the i-th critical band [13]. For critical bands less than 1000Hz the cut-off frequency of the LPF is set to $\frac{1}{8}$ of the critical bandwidth and for above 1000Hz the cut-off frequency is set to $\frac{1}{15}$ of the critical bandwidth. These fractions are selected to ensure that the envelope follows the signal closely but at the same time does not change too rapidly over time. The modulation products caused by a rapidly changing gain results in audible musical noise artifacts.

3.3 Perceptual Speech Enhancement

As mentioned in the previous sections, the signal is mapped to its corresponding loudness through the envelope of the signal in each critical band. Hence, to reduce the background noise in the signal, we operate on the envelopes of each critical band. Each critical band $c_i[n]$ is operated on independently. We drop the subscript i from all the equations here on for convenience, but recall the equations are computed for each critical band.

We assume the noise floor in each critical band corresponds to the minimum of the envelope in that critical band $e_{\min}$. We manipulate the dynamic range of the envelope such that the the noise floor is mapped to a fraction of its original value, which can be written as

$$\hat{e}_{\min} = K e_{\min}, \tag{11}$$

where K is an expansion factor, and $\hat{e}_{\min}$ is the modified noise floor. If $K < 1$, the noise floor is lowered. The dynamic range of the envelope is expanded non-linearly, so even if the estimate of the noise floor is incorrect the noise floor will be lowered. In this case, the noise suppression may not be as aggressive if the estimate was correct. Hence, with this type of processing the estimate of the noise floor does not have to be very accurate. We are also interested in not altering the speech envelope, in other words the quality of the speech. We can assume that the maximum of the envelope $e_{\max}$ corresponds to the speech level. Again, the dynamic range expansion is such that

$$\hat{e}_{\max} = e_{\max}. \tag{12}$$

This mapping of the dynamic range of the signal is shown pictorially in Figure 7.

To obtain noise suppression, we set $K < 1$ so that the dynamic range of the signal in each critical band is expanded. This is in contrast to the dynamic range compression to compensate for hearing loss, which is described in [13]. We modify the envelope according to equation (4), which is rewritten here in the discrete form

$$\hat{e}[n] = \beta e^{\alpha}[n], \tag{13}$$

where $e[n]$ is the envelope of the i-th critical band, α and β are parameters of the envelope

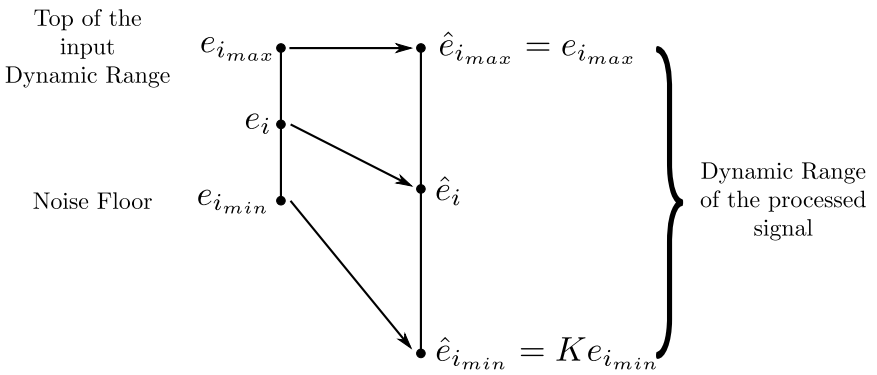


Figure 7: Dynamic mapping of the envelope for noise suppression.

expansion and $\hat{e}[n]$ is the modified envelope[1]. The power-law compression can be rewritten as a multiplicative gain,

$$\hat{e}[n] = G[n]e[n], \tag{14}$$

where $G[n] = \beta e^{\alpha-1}[n]$.

Taking the logarithm of Equation (13), we obtain

$$\log \hat{e}[n] = \alpha \log e[n] + \log \beta. \tag{15}$$

The parameters α and β are computed based on the constraints established in (11) and (12). Using Equations (11) and (12) in Equation (15) and solving for α and β, we obtain

$$\beta = e_{\max}^{(1-\alpha)} \tag{16}$$

and

$$\alpha = 1 - \frac{\log K}{\log M}, \tag{17}$$

where K is given by Equation (11), and

$$M = \frac{e_{\max}}{e_{\min}}. \tag{18}$$

The ratio in Equation (18) is proportional to the peak SNR in the i-th band. This ratio gives us an idea of the effective dynamic range of the input signal. The gain function multiplying

[1]We have expressed Equation (13) this way for convenience but in practice it is a good idea to normalize the envelope prior to the exponent for numerical reasons.

the sub-band signal is given by

$$\begin{aligned} G[n] &= \beta e^{(\alpha-1)}[n] \\ &= \left(\frac{e_{\max}}{e[n]}\right)^{P}, \end{aligned} \tag{19}$$

where $P = \frac{\log K}{\log M}$. Since $M \geq 1$, we have

$$G \begin{cases} \geq 1 & \text{when } K \geq 1 \\ < 1 & \text{when } K < 1. \end{cases}$$

3.4 Determining the Amount of Dynamic Range Expansion

As discussed in the previous section, if

$$0 < K < 1, \tag{20}$$

the envelope of the signal expands. We can rewrite Equation (19) as

$$G[n] = \left(\frac{e[n]}{e_{\max}}\right)^{-\frac{\log K}{\log M}}. \tag{21}$$

From Equation (20), it follows that $\log K < 0$. Therefore, Equation (21) can be rewritten as

$$G[n] = \left(\frac{e[n]}{e_{\max}}\right)^{\frac{|\log K|}{\log M}}. \tag{22}$$

If the value of $e[n]$ is close to $e_{\max}$, the instantaneous SNR is high. In this case, the value of K should be closer to 1 so that the gain is close to unity. On the other hand, if $e[n]$ is much less than $e_{\max}$, the instantaneous SNR is low and hence the value of K should be closer to 0 so that the gain G is small. One approach to obtain these values of K is to set

$$K[n] = \frac{e[n]}{e_{\max}}. \tag{23}$$

The gain G, obtained by using this form of K, is shown in Figure 8 for different values of the effective dynamic range.

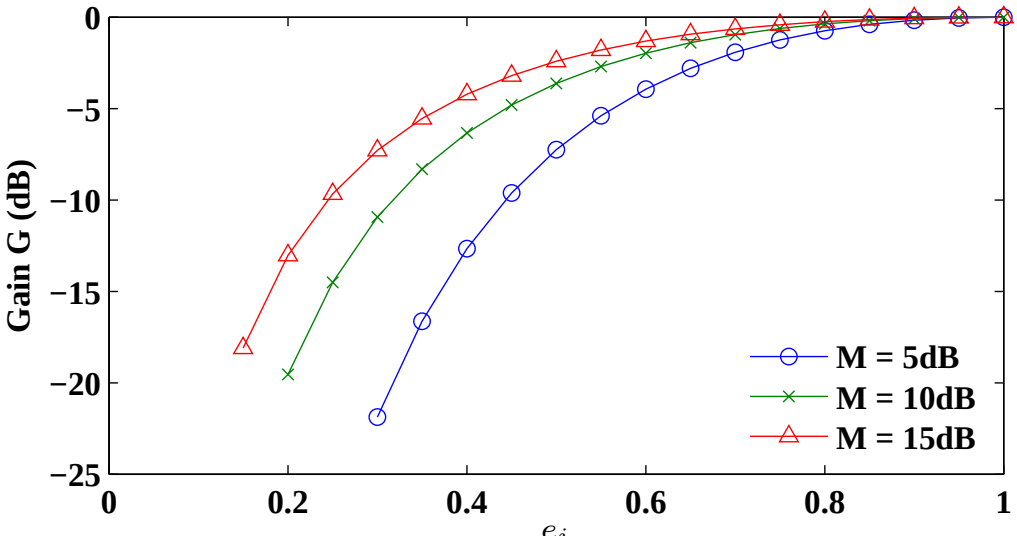


Figure 8: Graph of gain G vs. K for different values of the effective dynamic range.

By rewriting the expression for K, we find

$$\begin{aligned} K[n] &= \frac{e[n]}{e_{\max}} \\ &= \frac{e[n]}{e_{\min}} \cdot \frac{e_{\min}}{e_{\max}} \\ &= \frac{\mathrm{SNR}}{M}. \end{aligned} \tag{24}$$

The parameter K set in this form is proportional to the instantaneous normalized SNR.

3.5 Implementation

A block diagram, summarizing the implementation of the algorithm, is shown in Figure 9. The signal is analyzed as described in Section 3.2. We calculate the maximum of the envelope in each subband, which is the estimate of the signal level and the minimum of envelope, which is the estimate of the noise floor. For each sub-band the gain parameters K, β and α are calculated from Equations (23), (16) and (17), respectively. From Equation (19), the gain G is calculated and then the corresponding sub-band is multiplied by this gain. The sub-bands are then added to obtain the noise-suppressed signal. Since the envelope is varying more slowly than the signal, computational requirements may be relaxed by calculating the gain at a slower rate commensurate with the envelope bandwidth.

As a result of the low complexity of this algorithm, it can be easily implemented in real

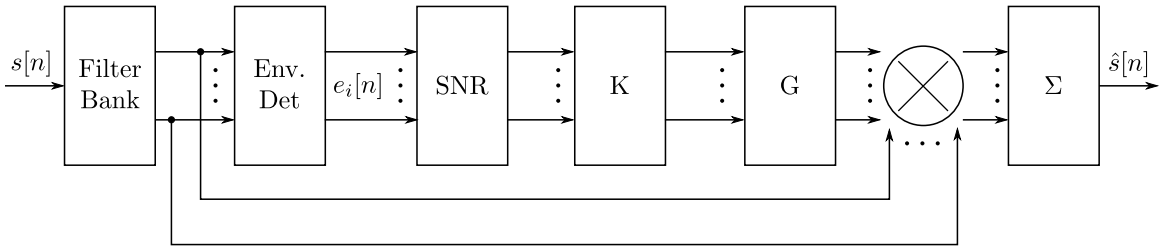


Figure 9: Block diagram of AGC-based noise-suppression algorithm.

time. For real-time implementation the signal may be processed in blocks. The block size can be determined based on the memory available. Care needs to be taken during block processing to maintain continuity between the blocks since a discontinuous gain between blocks can cause undesirable artifacts in the output. The filter states need to be preserved from the previous block and used for the processing of the current block. The peak SNR calculated in Equation (18) is the peak SNR of each critical band and not the peak SNR of each block. Hence, the signal level estimated by $e_{\max}$ is the maximum of the entire critical band and not just a single block. This maximum can be calculated as

$$(e_j)_{\max} = \max((e_j)_{\max}, \gamma(e_{j-1})_{\max}), \tag{25}$$

where $\gamma \approx 1$ but $\gamma < 1$ and $(e_j)_{\max}$ is the maximum of the envelope of the current j-th block of the i-th critical band of the signal and $(e_{j-1})_{\max}$ is the maximum of the previous $(j-1)$-th block of the i-th critical band. The gain continuity can be obtained by interpolating the gain at the end of the previous block to the gain in the current block. Again, recall each of these equations are calculated for each critical band, and the subscript i indicating the critical band is dropped for convenience.

3.6 Results

Figures 10 to 12 show the spectrograms of the original noisy signal and the output of the AGC noise-suppression algorithm. It is clear from the spectrograms that the noise level is reduced without distorting the speech spectrum.

In signals with very low SNR (approaching 0 dB), the approximation of the SNR is not accurate since the minimum of the envelope may not correspond to the noise floor. In this case, the rapidly changing gain modulates the noise in the higher frequency bands,

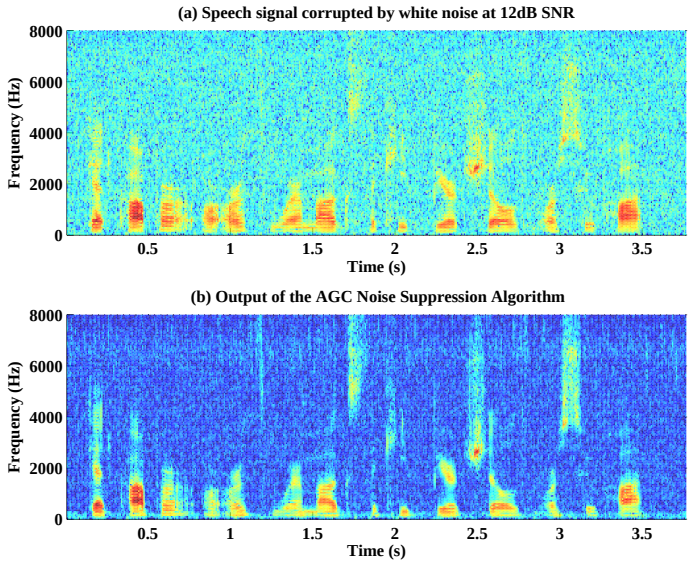


Figure 10: Results of AGC-based perceptual speech-enhancement algorithm for white noise corrupted speech at 12 dB SNR. The upper panel is the noisy-speech signal and and the lower panel is output of the AGC-based noise-suppression algorithm.

resulting in audible musical noise artifacts. This can be seen in the spectrogram of the signal in Figure 12. Yet the quality of the speech is preserved. This is validated by the subjective testing, which is described next.

3.7 Subjective Testing

A subjective test was conducted to evaluate the performance of our algorithm compared to three other standard noise-suppression methods – spectral subtraction (SpecSub) [9], multi-band spectral subtraction(Mband) [27], and an iterative Wiener algorithm based on all-pole speech production model (Wiener) [32]. The code for these algorithms was obtained from [33]. The algorithms were tested in four different noisy conditions and at three different noise levels. The noise samples were obtained from the NoiseX database. The noisy speech samples were generated by adding white noise, babble noise, F-16 cockpit noise, and the noise inside a military vehicle (Leopard 1) at 5 dB, 12 dB, and 20 dB SNR.

Eleven native English speaking subjects were presented with pairs of speech samples processed with different noise-suppression algorithms and were asked to rate the quality of one sample compared to the other. The subjects were asked to rate the quality of the speech (Q) based on how natural the sample was, speech intelligibility, and distortions present in

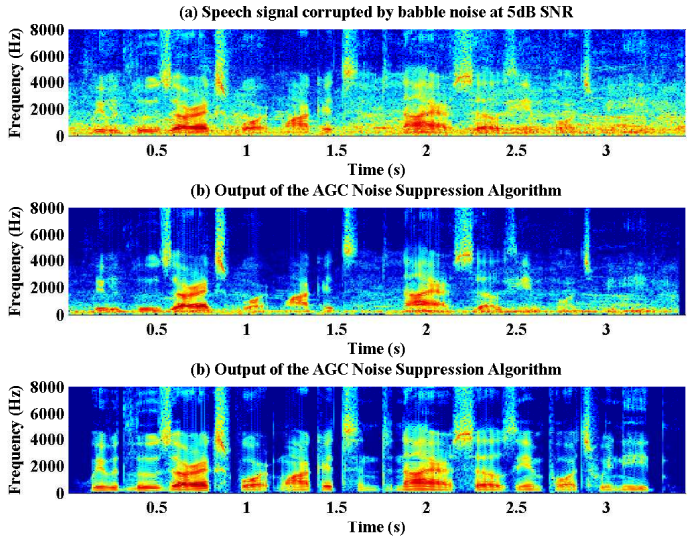


Figure 11: Results of AGC-based perceptual speech-enhancement algorithm for babble noise corrupted speech at 5 dB SNR. The upper panel is the noisy-speech signal and and the lower panel is output of the AGC-based noise-suppression algorithm.

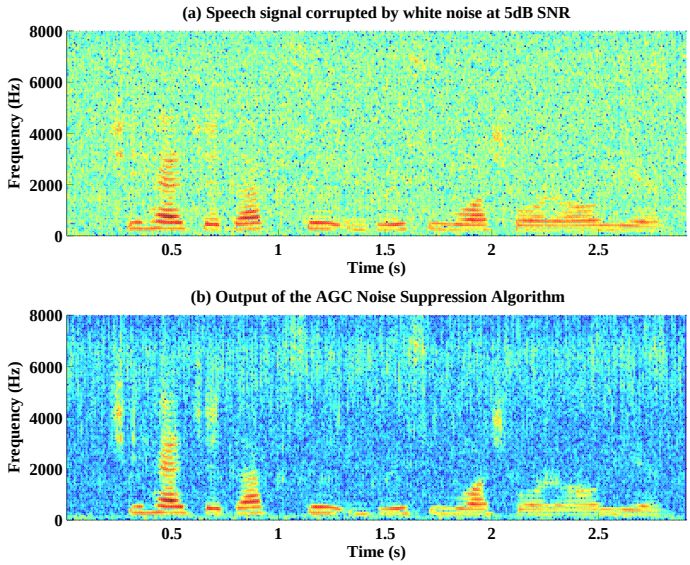


Figure 12: Results of AGC-based perceptual-speech enhancement algorithm for white noise corrupted speech at 5 dB SNR. The upper panel is the noisy-speech signal and the lower panel is output of the AGC-based noise-suppression algorithm.

the sample. The allowable responses were that the second sample was much better (3), better (2), slightly better (1), about the same (0), slightly worse (-1), worse (-2), or much worse (-3) than the first sample. They were also asked to rate the overall noise level (N) of one sample compared to the other. The three possible ratings in this case were: the second sample is less noisy (1), about the same (0), or more noisy (-1) than the first sample. The

subjects were allowed to replay the samples as many times as they liked. 36 pairs of samples were presented to each subject.

The results of the subjective test are summarized in Table 1-3. The ratings in the tables indicate on an average how the algorithms mentioned in the first column were rated compared to our AGC-based algorithm in terms of quality of speech (Q) and in terms of noise level (N). The values in the table correspond to the ratings described in the previous paragraph. Overall, we see that our algorithm outperformed standard methods in terms of preserving the quality of the speech. While it was rated at par in terms of the noise level in the processed output.

Table 1: Subjective test results comparing perceptual speech enhancement to other standard methods in terms of overall quality.

	Q	N
SpecSub	-1.15	-0.06
MBand	-0.86	-0.09
Wiener	-0.56	0.43

Table 2: Subjective test results comparing perceptual speech enhancement to other standard methods in terms of noise levels and quality of speech for different types of noise.

	White		Babble		F16		Leopard	
	Q	N	Q	N	Q	N	Q	N
SpecSub	-1.09	-0.15	-1.33	-0.12	-0.87	0.27	-1.30	-0.21
MBand	-0.81	-0.15	-1.39	-0.30	-0.78	0.18	-0.45	-0.09
Wiener	-0.27	0.54	-0.87	0.18	-0.87	0.45	-0.24	0.57

Table 3: Subjective test results comparing perceptual speech enhancement to other standard methods in terms of noise levels and quality of speech for different levels of noise.

	5dB SNR		12dB SNR		20dB SNR	
	Q	N	Q	N	Q	N
SpecSub	-1.45	0.30	-1.57	-0.03	-1.57	-0.54
MBand	-0.75	0.51	-0.93	0.00	-1.70	0.87
Wiener	-1.72	0.66	-0.36	0.96	-0.18	0.12

CHAPTER IV

PERCEPTUAL NOISE SUPPRESSION AS BLIND SOURCE SEPARATION POST PROCESSING

In general, single-channel noise-suppression algorithms rely on estimates of the noise spectrum from the given noisy speech data. Based on the estimated noise and speech spectrum, a noise-suppression gain is then applied to each frequency bin. These methods work well under stationary noise conditions since an accurate noise estimate can be obtained over time. In the presence of non-stationary noise, it becomes difficult to track and suppress time-varying noise [33]. Multi-channel noise-suppression methods have attempted to fill in this performance gap by recording the noisy speech at different spatial locations. Dual-microphone-based blind source separation (BSS) methods can be used to separate speech from additive noise using an adaptive linear filter. While the linear filter helps in retaining the speech quality and intelligibility of the desired talker, the amount of noise rejected is limited especially in the presence of diffused noise. The presence of diffused noise is a more realistic scenario. The capability of a BSS method is severely hampered because of limited degrees of freedom. However, the secondary channel of the BSS algorithm rejects the desired talker by creating a null towards the desired talker [52]. This secondary channel can be used to obtain a better estimate of the residual noise present in the primary channel.

In this chapter, we first describe the blind source separation algorithm we use to separate the signals before post processing. Then, we describe the experimental setup that was used to test the proposed algorithms in a more realistic scenario and our conclusion on the optimal location of the microphones on a cellphone to obtain maximum separation. We then discuss the proposed post-processing technique based on perceptual processing and its performance.

4.1 Info-Max Blind Source Separation

We use a stochastic gradient adaptive learning algorithm for BSS [23]. A block diagram of this method is shown in Figure 13. The feedback structure of the separation network ensures that the separating filters remove redundancies across channels rather than the redundancies within the channel. This prevents the separating filters from converging to whitening filters. Moreover, the direct path filters are constrained to scalers to also avoid the whitening of the sources.

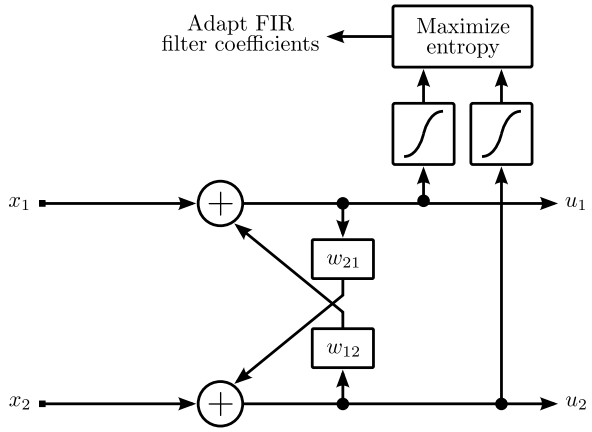


Figure 13: Constrained-BSS configuration. This approach assumes two sources, each of which is closer to a different microphone. The advantage of this configuration is that there is no source permutation ambiguity.

In this method, the signals are separated by minimizing the mutual information between the approximated cumulative density functions (CDF) of the separated sources. The CDF of the signal can be approximated by applying a non-linear function, such as the hyperbolic tangent (tanh), to the approximated output signal u_j. For speech signals, minimizing the mutual information is equivalent to maximizing the entropy of the signal. The entropy of the signal is given by

$$H(y_i) = -\mathrm{E}[\log(f_{y_i}(y_i))], \tag{26}$$

where $f_{y_i}(y_i)$ is the probability density function (PDF) of $y_i = \tanh(u_i)$. The PDF of the output can be written as

$$f_{y_i}(y_i) = \frac{f_{x_i}(x_i)}{\det(J)}, \tag{27}$$

where J is the Jacobian of the unmixing network. Maximizing the entropy leads to maximizing E[log(det(J))]. A stochastic gradient rule can be computed from this to obtain the unmixing filter updates, which is of the form

$$\Delta w_{ij}(k) \propto \hat{y}_i \; u_j(n-k), \tag{28}$$

where

$$\hat{y}_i = \frac{\partial}{\partial y_i}(\frac{\partial y_i}{\partial u_i}). \tag{29}$$

The separated sources u_i are obtained by applying these unmixing filters w_{ij} to the mixtures x_j

$$u_i[n] = x_i[n] + \sum_{k=0}^{L} w_{ij}[k] x_j[n-k]. \tag{30}$$

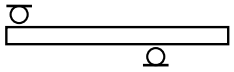
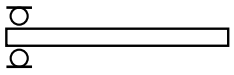
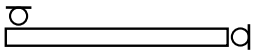
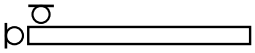
Since the separation is based on an adaptive FIR filter, this algorithm is ideal for hardware implementation. Moreover, this algorithm can separate sources that have been mixed in a convolutive environment.

4.2 Experimental and Recording Setup

To simulate a real environment, we carried out our experiments in a room of dimensions 12ft by 9ft 11in by 8ft 2in. Two omni-directional microphones were mounted on the cell phone. One microphone that served as the primary microphone of the the cell phone was placed at the bottom front center of the cell-phone. The secondary microphone was placed at different locations along the back and side of the cell phone to test the best location for the secondary microphone to obtain maximum source separation for cellphone applications. We tested five realistic microphone configurations listed in Table 4. The cell-phone user was simulated by placing a loudspeaker near the primary microphone whereas the interfering source was placed at different locations in the room. Experiments with diffused noise were carried out in a slightly bigger room of dimensions 10ft 2in by 12ft 3in by 10ft. The diffused noise was simulated by placing an interfering speaker in the corner of the room facing the wall.

Speech was played on the primary speaker and noise or speech from an interfering speaker was played on the secondary speaker. These signals were recorded by the two

Table 4: Microphone configurations tested in the BSS experiments to find the optimal microphone positions.

Config	Microphone Placement
M1	
M2	
M3	
M4	
M5	

microphones, and the data was captured using an acquisition board, while the data was stored and processed on a PC. Nine different test cases were considered with different combinations of the primary speaker and the interference. These cases are listed in Table 5. The location of the primary speaker was fixed whereas the interfering source was placed at different locations in each of the test cases. In cases S5–S7 the primary speech was characterized by pauses in the conversation.

4.3 Performance Assessment

Signal-to-interference ratio (SIR) is the conventional measure to assess the performance of BSS [42]. However, the computation of SIR requires a priori knowledge of the mixing as well as the unmixing filters to determine the signal and interference contributions. In a real recording environment the mixing filters cannot be measured, hence we cannot rely on SIR for performance evaluation. Instead, we use a simple SNR measure to evaluate the performance of the algorithm. We calculate the SNR before the BSS processing and compared it to the SNR after the processing. The SNR can be calculated by determining the noise energy during the silence periods of the speech. The exact location of these silence intervals can be determined from the clean speech that is played through the primary loudspeaker. Note that this information about the clean speech and noise is not available

Table 5: Test cases used in the BSS experiments.

Test Case	Primary Speaker	Secondary/Interfering Noise
S1	Female	Speech
S2	Female	Train
S3	Female	Pub
S4	Male	Car
S5	Male and Female with pauses	Train station
S6	Female with pauses	Pub
S7	Male with pauses	Train
S8	Female	Diffused pub noise
S9	Female	Diffused pub noise and male speaker

in practice and is used here only for the sake of performance evaluation.

4.4 Impact of Microphone Positions on Speech Separation

Initially, we evaluated the BSS performance using five microphone configurations in the five test cases M1–M5 listed in Table 5. We used unmixing filter length of $P = 128$ in our experiments. We eliminated microphone configuration M4 earlier in our investigation for further consideration because of its relative poor performance in the first five test cases of Table 5. For the four remaining microphone configurations, we made recordings at five different levels of input SNR – 15 dB, 10 dB, 5 dB, 0 dB, and < 0 dB. The results for the test cases S6 and S7 are shown in Figure 14 and Figure 15 respectively. From both figures it is clear that microphone configuration M1 consistently results in better performance over a range of input SNR values.

We also evaluated the noise suppression performance using different unmixing filter lengths. Figure 16 shows the performance of the BSS algorithm for three different filter lengths of 64, 128, and 256 using microphone configuration M1. We note that for high levels of input SNR, the output SNR is somewhat the same for the three filter lengths. While, for low levels of input SNR, the performance improves with increasing filter length.

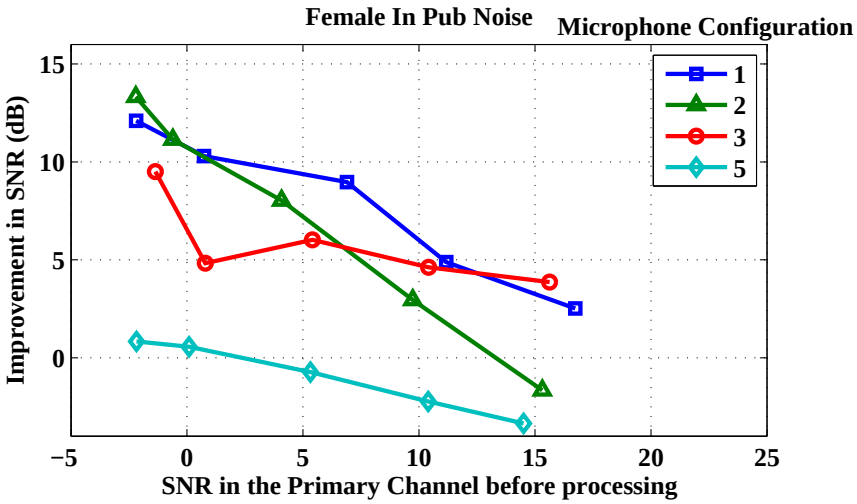


Figure 14: SNR improvement for different input SNR for test case S6 using different microphone configurations. Unmixing filter length of $P = 128$ was used.

However, the processing complexity increases with the filter length. To ease this complexity-performance trade-off, we suggest to use the length of $P = 128$.

4.5 Post Processing using a Wiener Filter

As mentioned earlier, the source separation is not perfect. We can assume a signal model in which the the BSS output can be expressed as the sum of the ideal signal and an interference signal consisting primarily of the signal isolated in the other BSS output:

$$y_1[n] = s_1[n] + \gamma y_2[n],$$

where $s_1[n]$ represents the clean speech for channel 1, $y_1[n]$ represents the BSS output for channel 1 and γ is a residual mixing gain constant. It is likely that $s_1[n]$ is also corrupted by other added noise and distortion sources but, in our experience, these are typically much lower in amplitude than the amplitude of $\gamma y_2[n]$.

An algorithm based on using a Wiener Filter for post processing is described in [37]. The output of the secondary channel of BSS gives us an estimate of the residual-noise spectrum present in the primary speech. This estimate can be used to drive a Wiener filter. The equation for the Wiener filter in the frequency domain can be written as

$$H(\omega) = \frac{\hat{P}_{y_1} - \gamma^2 \hat{P}_{y_2}}{\hat{P}_{y_1}}, \tag{31}$$

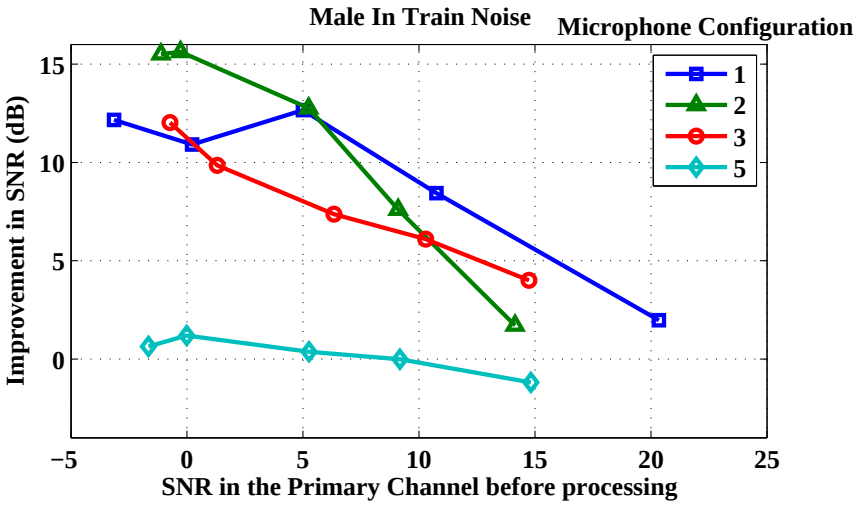


Figure 15: SNR improvement for different input SNR for test case S7 using different microphone configurations. Unmixing filter length of $P = 128$ was used.

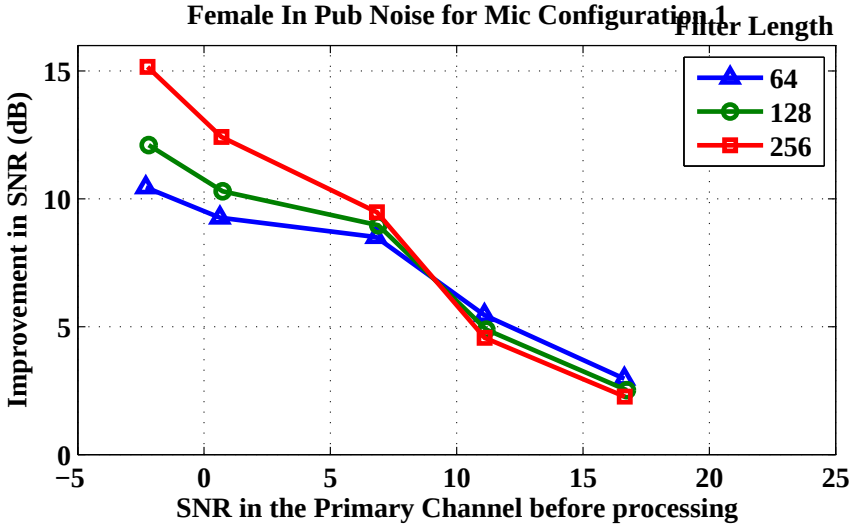


Figure 16: SNR improvement for different input SNR for test case S6 using different unmixing filter lengths P. Microphone configuration M1 was used.

where $\hat{P}_{y_1}(\omega)$ and $\hat{P}_{y_2}(\omega)$ are estimates of the power spectral densities of $y_1(n)$ and $y_2(n)$, respectively.

In earlier work by Park *et al.* a similar function is proposed, but the scaling factor in front of $\hat{P}_{y_2}$ is not present [41]. It is likely that this system would enhance the output without the scaling factor γ in much the same way that the over-subtraction system would in the spectral-subtraction step in most Wiener filter implementations. However, this performance would be dependent on the particular BSS implementation and environment.

Noohi *et al.* demonstrate an improvement in SIR using mixture signals generated using a room simulation tool [37]. They use a robust technique for estimating γ for each BSS output in a multiple-microphone configuration. However, no listening tests or other speech quality measurements were reported.

4.6 Results of Wiener Filter Post Processing

Post processing is performed on sub-band signals at the output of the BSS block as shown in Fig. 17. For the Wiener filter, the frequency decomposition is performed using the short-time Fourier transform generated using overlapping, windowed FFTs.

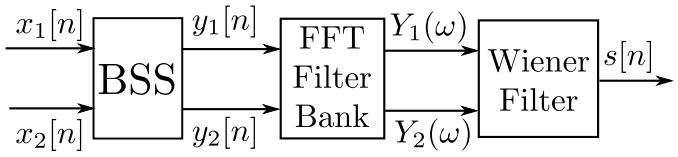


Figure 17: Post processing is performed using an FFT filter bank for the adaptive Wiener filtering.

We processed the recordings that were made by the setup described in Section 4.2 using the Wiener filter described in Section 4.5. We conducted a subjective test to evaluate the performance of this algorithm. Seven native English speaking subjects were recruited. They were asked to rate the quality of speech, noise level and overall signal on a scale of 1-to-5; 1 being the worst and 5 being the best. They were presented with examples of a clean speech and a noisy speech that were obtained from the microphone recordings. The subjects were allowed to replay the samples as many times as they liked. Forty speech samples were presented to each subjects. The samples presented during the test included sixteen samples that were the primary output of BSS, sixteen samples that were the output of the Wiener filter post processing (WF-PP), six mixtures obtained from the microphones and two clean speech samples. These samples were presented to the subjects in a random order and no information about the nature of each sample was given to the subjects. The results are presented in Table 6.

As we mentioned earlier, even though an algorithm can achieve an improvement in SIR (signal-to-interference ratio) it is not necessary that the sound quality is preserved. In our

Table 6: Results of the subjective test to determine the ratings of the Wiener filter post processing. The average rating of each of the mentioned speech samples is presented. The ratings are on a scale of 1-to-5, 1 being the worst and 5 being the best.

	Speech Quality	Noise Level	Overall
WF-PP	3.6	4.1	3.4
BSS	4.1	3.2	3.6
Mixture	2.9	2.8	1.9
Clean	4.8	4.6	4.8

tests, the Wiener filter post-processed signals were judged by listeners to have lower residual noise than the BSS output signals. However, the overall quality was judged to be lower. This is because the post-processing filter introduces slight distortions to the speech and the residual noise is no longer simply another talker, which sounds somewhat natural, but a signal that has some musical and artificial qualities. In the following sections, we propose an algorithm that not only suppresses the noise but also preserves the speech quality.

4.7 Perceptual Post Processing

Instead of blindly reducing the amount of noise in a speech signal, which may introduce artifacts into the speech, we propose to suppress the noise based on the human perceptual auditory model. One such method for a single-microphone case was described in Section 3.3. Noise suppression is obtained by mapping the minimum of the envelope in each critical band that corresponds to the noise floor to a fraction of its value. Since this mapping is based on the human perceptual auditory model, the resulting speech sounds natural. In this section we demonstrate that this method can be easily extended to a dual microphone case.

Given that the signal of interest is contained in the channel $y_1[n]$, which has been obtained from BSS, we refer to this as the primary channel. Additionally, we refer to the channel $y_2[n]$ as the secondary channel. A block diagram of our proposed perceptual post-processing method is show in Figure 18. The output obtained from the BSS processing is applied to a filter bank to decompose the signal into sub-bands. A constant-Q filter bank is used. We then extract the envelope from each sub-band and estimate the SNR in each

sub-band. The gain G that is applied to the sub-bands is calculated using this estimate of the SNR. The weighted sub-bands are then added to obtain the noise-suppressed speech. More information about the constant-Q filter bank and the envelope detector can be found in [40].

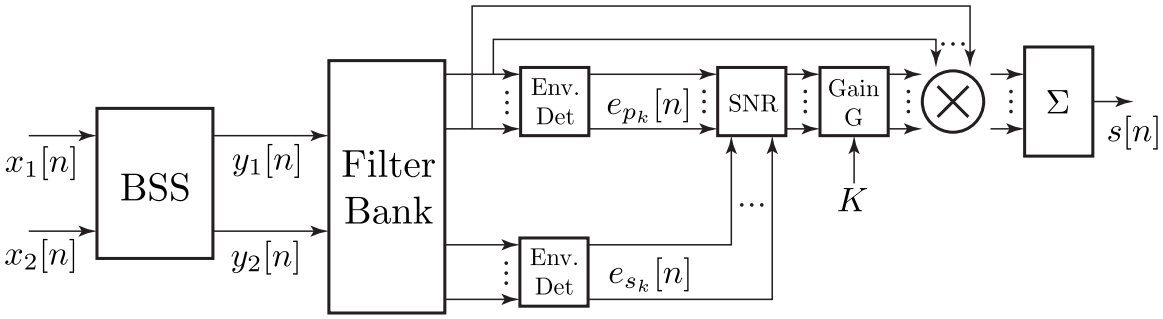


Figure 18: Post processing is performed using a constant-Q filter bank for the perceptual speech-enhancement algorithm.

The gain G is calculated as

$$G = \beta(e_{p_k}[n])^{\alpha-1}, \tag{32}$$

where $e_{p_k}[n]$ is the envelope of the k-th frequency band of the primary channel and β and α are scaling and expansion factors that can be calculated on the basis of the SNR of the signal (M) and the amount of expansion (K) that is desired. These factors are calculated as

$$\beta = (\max(e_{p_k}[n]))^{(1-\alpha)}, \tag{33}$$

$$\alpha = 1 - \frac{\log K}{\log M}. \tag{34}$$

The envelope of the primary speech gives us an estimate of the speech level, while the secondary channel scaled by the residual mixing gain $\gamma[n]$ gives us an idea of the noise level present in the primary signal. The average SNR can be estimated by

$$M = \frac{\max(e_{p_k}[n])}{\max(\gamma[n] \cdot e_{s_k}[n])} \tag{35}$$

where $\max(e_{p_k}[n])$ and $\max(e_{s_k}[n])$ are the maximum of the envelopes of the k-th frequency band of the primary and secondary channel, respectively.

Since we have access to the entire envelope of the primary and secondary signals we can determine the value of $\gamma[n]$. When the primary speech is not active we can find $\gamma[n]$ by the following equation

$$\gamma[n] = \frac{e_{p_k}^2[n]}{e_{s_k}^2[n]} \tag{36}$$

When the speech is active, we set the value of $\gamma[n]$ to the mean of the $\gamma[n]$ calculated during the silence period. The combination of the fact that we have access to an accurate estimate of the noise spectrum and a time-varying $\gamma[n]$ allows us to handle non-stationary noise cases.

The value of K, which determines how much the envelope expands, is set to a value of 0.03. At this value, the processing achieved maximum noise suppression without audible distortion. Using (33) and (34) in (32), the gain G can be calculated. This gain is then applied to each sub-band and all such sub-bands are then added up to obtain the noise-suppressed speech.

4.8 Results

Figure 19 and 20 show the spectrograms of the post-processed BSS, the output of BSS and the actual mixture. It is clear from these spectrograms that the noise level has been reduced without distorting the speech spectrum.

A listening test was conducted to determine the subjective quality of the post-processed BSS signals using the perceptual post processing. Ten native English speakers were recruited. They were asked to rate the speech samples presented to them on the basis described in Section 4.6. Forty samples were presented to the subjects. These samples included ten samples that were the unprocessed outputs of BSS, ten samples that were perceptually post processed (P-PP) using the algorithm we propose, ten samples that were post processed using a Wiener Filter (WF-PP), five mixtures obtained from the microphones and five clean speech samples.

From Table 7, we can see that the proposed perceptual post-processing does not alter the speech quality of the output of BSS. There is also a dramatic improvement in the noise level and overall rating as compared to the unprocessed output of BSS and the post processing

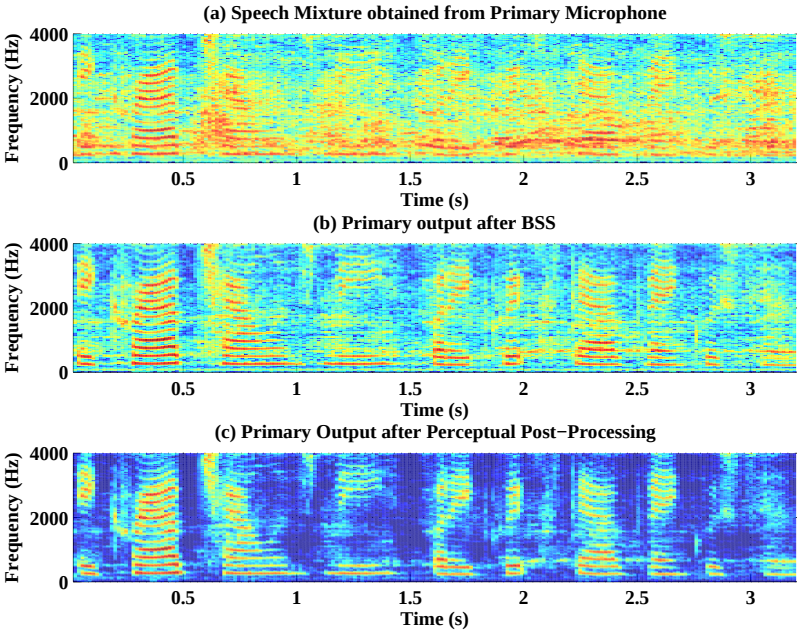


Figure 19: Spectrogram of the mixture, output of BSS and output of perceptual post processing. The SNR of the mixture is about -2 dB.

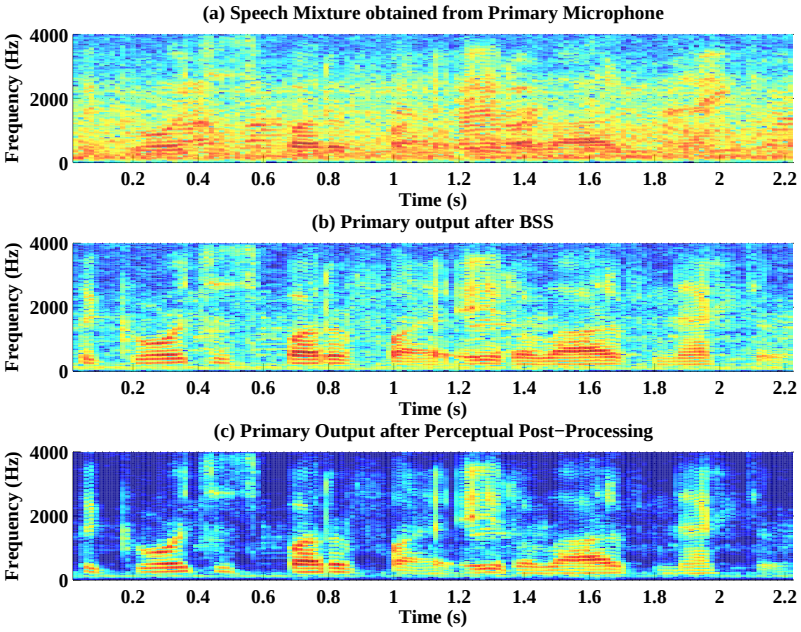


Figure 20: Spectrogram of the mixture, output of BSS and output of perceptual post processing. The SNR of the mixture is about 0 dB.

done using a Wiener filter.

Table 7: Results of the subjective test to determine the ratings of the perceptual post processing. The average rating of each of the mentioned speech samples is listed.

	Speech Quality	Noise Level	Overall
P-PP	3.9	3.9	3.8
WF-PP	3.1	3.2	2.9
BSS	3.9	2.7	3.2
Mixture	2.7	1.3	1.9
Clean	4.9	4.9	4.9

CHAPTER V

IMPLICIT GAIN SMOOTHING

In this chapter, we present a more realistic implementation of a psychoacoustic-motivated dual-microphone noise-suppression system, making the algorithm more suitable for a real-time implementation. The original algorithm uses a dual-microphone blind source separation algorithm as the front end, which is followed by perceptual post processing that is based on the human perceptual auditory system. However, in the algorithm described in Chapter 4, the noise-suppression parameters were set based on the prior knowledge of the signal, which is not available in a causal implementation. In this paper, we will show why updating the noise-suppression parameters continuously over time generates musical-noise artifacts which were not seen in the original non-causal implementation. We will then propose an implicit gain smoothing technique based on the Ephraim-Malah suppression rule, which reduces the musical noise artifacts.

5.1 Causal Implementation

Figure 21 summarizes the two-microphone noise-suppression algorithm that is described in Chapter 4. Mixtures $x_1[n]$ and $x_2[n]$ are captured by two microphones. These mixtures are unmixed using an info-max BSS algorithm, details of which can be found in [39] and Chapter 4, to obtain separated sources $y_1[n]$ and $y_2[n]$ respectively. The primary channel $y_1[n]$ contains the desired talker and the residual diffused noise. While the secondary channel $y_2[n]$ contains the separated diffused noise with the desired talker sufficiently suppressed.

The primary and secondary channels, $y_1[n]$ and $y_2[n]$, are decomposed into frequency bands using a filter bank that models the cochlear filter bank. We then extract the envelope $e_{p_k}[n]$ and $e_{s_k}[n]$ from the k-th subband of the primary and secondary channels respectively. Details of the filter bank and the envelope extraction can be found in 3.2. We assume that the maximum of the envelope in each subband of the primary channel corresponds to the

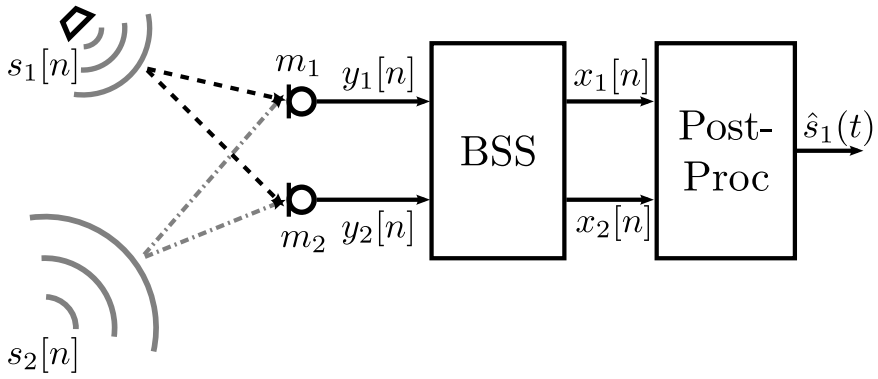


Figure 21: Basic block digram of the multi-channel noise-suppression algorithm.

speech level $e_{k_{\max}}$ in the corresponding subband. The speech level is updated every 1 ms as

$$e_{k_{\max}} = \max(\alpha_{\text{speech}} e_{k_{\max}}, e_{p_k}[n]), \tag{37}$$

where α_{speech} is a forgetting factor which is set such that the time constant is 4 times the time constant of the LPF used to extract the envelope. This value of the time constant is selected so that the estimate of $e_{k_{\max}}$ quickly tracks the peaks of the envelope but does not fall too rapidly once the signal level goes down. The noise level $e_{k_{\min}}$ is estimated every 1 ms as

$$e_{k_{\min}} = \alpha_{\text{noise}} e_{k_{\min}} + (1 - \alpha_{\text{noise}}) e_{s_k}[n], \tag{38}$$

where $\alpha_{\text{noise}} = 0.95$. This value of α_{noise} tracks the noise level closely while smoothing out any rapid changes in the noise level. The dynamic range, in other words the peak SNR in each subband can be calculated as $M_k = \frac{e_{k_{\max}}}{e_{k_{\min}}}$.

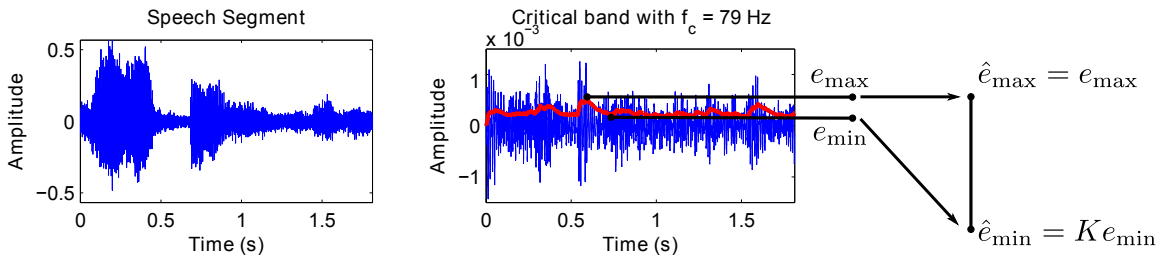
The noise-suppression gain G is calculated as

$$G_k[n] = \beta_k (e_{p_k}[n])^{\alpha - 1}, \tag{39}$$

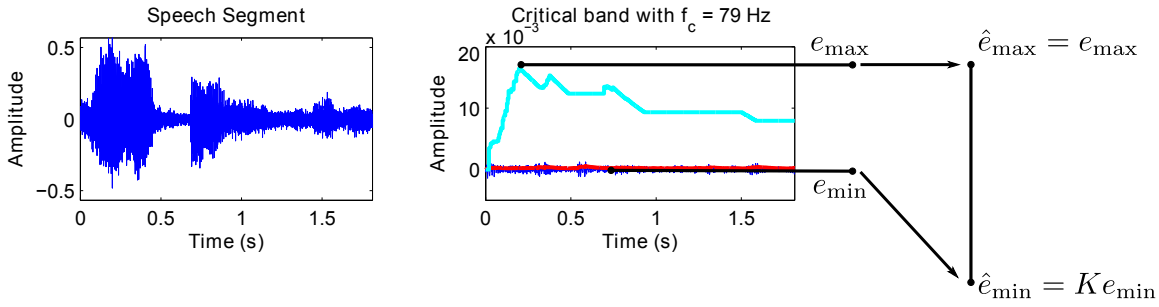
where $\beta = (e_{k_{\max}})^{(1-\alpha)}$ and $\alpha = 1 - \frac{\log K}{\log M}$. K is the expansion factor which determines how much the noise floor $e_{k_{\min}}$ is suppressed. As described in Chapter 4, initially the value of K is set to 0.01.

5.2 Analysis of Artifact generation

Two main types of artifacts are observed in our implementation. One that is generated in frequency bands that have very-low SNR and the other that is a result by the causal



(a) Cause of low-frequency artifacts



(b) Remedy of low-frequency artifacts

Figure 22: Low-frequency artifacts

implementation of the algorithm. In this section we will investigate the sources of the artifacts and propose techniques that will reduce these artifacts.

In the low-frequency bands (below 200 Hz), where the noise may dominate the speech information present, if the speech level $e_{k_{\max}}$ is determined using (37) then $e_{k_{\max}}$ may still correspond to the noise level. If we apply a non-linear expansion such that this estimated speech level in the low-frequency bands remains constant, then the noise-suppression gain will cause unnatural modulations to these bands where the noise is dominant. This phenomenon is shown in Figure 22(a). Such unnatural expansion will cause audible low-frequency artifacts. Instead, if we translate the average speech level from the frequency bands where the speech is dominant to the low-frequency bands, which is shown in Figure 22(b), then the entire band will be suppressed, hence reducing the amount of audible artifacts.

To suppress noise, a time-varying spectral gain is applied to the noisy signal. This time-varying noise-suppression gain modulates the noisy signal and in this process generates modulation artifacts that may be heard as musical noise. The rate of change of the gain controls the balance between noise suppression and the musical noise. A slower rate of gain change will not be able to track and suppress fast changes in the signal level. On the other

hand, if the rate of gain change is too high, then the modulation artifacts are heard as musical noise [5].

In the non-causal implementation of the algorithm described in [39], the long-term dynamic range M_k and the expansion factor K does not vary with time. In this case, the rate of change of the gain G with respect to time is given by,

$$\frac{dG_k}{dt} = -\frac{\log K}{\log M_k} \cdot \frac{G_k}{e_k} \cdot \frac{de_k}{dt}. \tag{40}$$

The rate of change of the envelope is limited by the cut-off frequency of the low-pass filter used to extract the envelopes in each subband. The cut-off frequencies of the low-pass filter that have been used to extract the envelopes is such that the rate of change of the gain does not cause musical noise. However, if the cut-off frequency is increased, in other words, if the rate of change of the envelope is increased, then the gain calculated as per the non-causal implementation of the algorithm will also generate musical noise.

In the causal implementation, if the estimate of the long-term dynamic range M_k is updated over time, then the rate of change of the gain with respect to time increases to

$$\begin{aligned} \frac{dG_k}{dt} = &- \frac{\log K}{\log M_k} \cdot \frac{G_k}{e_k} \cdot \frac{de_k}{dt} \\ &+ G_k \cdot \ln\left(\frac{e_k}{e_{k_{\max}}}\right) \cdot \frac{d}{dt}\left(-\frac{\log K}{\log M_k}\right). \end{aligned} \tag{41}$$

The second term in (41) causes an increase in $\frac{dG_k}{dt}$, which results in annoying musical noise artifacts. Hence, we need to reduce the rate of change of the gain to eliminate the musical noise. The cut-off frequency of the low-pass filter can be reduced so that the rate of change of the envelope is reduced. However, a slower envelope will not track the speech and noise accurately and will attenuate some of the speech cues. Another common technique to reduce the rate of change of the gain is to smooth the gain variations over time. Typically a single-pole low-pass filter is used to smooth the gain. However, this explicit reduction of the rate of change of the gain introduces temporal smearing of the spectrum, which leads to additional echo-type artifacts. In the next section, we will explain a more implicit gain-smoothing technique which is inspired by the Ephraim-Malah suppression rule [18].

5.3 Motivation for Implicit Gain Smoothing

The Ephraim-Malah suppression rule (EMSR) [18] achieves a reduction of the rate of change of the gain in a rather implicit manner. The EMSR is a minimum mean-square error (MMSE) short-time spectral-amplitude estimator. The gain [12] is estimated using

$$G^{\mathrm{EM}} = \frac{\sqrt{\pi}}{2}\sqrt{\left(\frac{1}{1+\mathcal{R}^{\mathrm{EM}}_{\mathrm{post}}}\right)\left(\frac{\mathcal{R}^{\mathrm{EM}}_{\mathrm{prio}}}{1+\mathcal{R}^{\mathrm{EM}}_{\mathrm{prio}}}\right)}\times \mathbf{M}\left[\left(1+\mathcal{R}^{\mathrm{EM}}_{\mathrm{post}}\right)\left(\frac{\mathcal{R}^{\mathrm{EM}}_{\mathrm{prio}}}{1+\mathcal{R}^{\mathrm{EM}}_{\mathrm{prio}}}\right)\right], \tag{42}$$

where $\mathcal{R}^{\mathrm{EM}}_{\mathrm{post}}$ is the *a posteriori* SNR and $\mathcal{R}^{\mathrm{EM}}_{\mathrm{post}}$ is the *a priori* SNR, both of which are evaluated in each short-time frame for all spectral-frequency bins. The function $\mathbf{M}[\cdot]$ is defined by

$$\mathbf{M}[\theta] = \exp\left(-\frac{\theta}{2}\right)\left[(1+\theta)\, I_0\left(\frac{\theta}{2}\right) + \theta I_1\left(\frac{\theta}{2}\right)\right], \tag{43}$$

where I_0 and I_1 are the modified Bessel functions of zeroth and first order, respectively [18].

The *a posteriori* SNR, $\mathcal{R}^{\mathrm{EM}}_{\mathrm{post}}$, is defined by

$$\mathcal{R}^{\mathrm{EM}}_{\mathrm{post}}(i,k) = \frac{|X(i,k)|^2}{\mu_k} - 1, \tag{44}$$

where $X(i,k)$ is the k-th frequency bin of short-time Fourier transform of the i-th time frame of the noisy speech and μ_k is the estimate of the noise power in the k-th frequency bin. The *a priori* SNR $\mathcal{R}^{\mathrm{EM}}_{\mathrm{prio}}$, is defined by

$$\mathcal{R}^{\mathrm{EM}}_{\mathrm{prio}}(i,k) = \frac{|S(i,k)|^2}{|N(i,k)|^2}, \tag{45}$$

where $S(i,k)$ and $N(i,k)$ are the k-th frequency bin of short-time Fourier transform of the i-th time frame of the clean speech and noise, respectively. Since, $\mathcal{R}_{\mathrm{prio}}$ cannot be found exactly it can be estimated using a directed-decision approach using the processed noise-suppressed signal from the previous frame [18].

From (42), we can see that the EMSR gain is a function of both the *a priori* SNR and *a posteriori* SNR. Figure 24(a) shows how the EMSR gain varies with the *a priori* SNR for different values of the *a posteriori* SNR. From this figure we see that in the event of a disagreement between the *a priori* and *a posteriori* SNR at lower values of the *a priori*

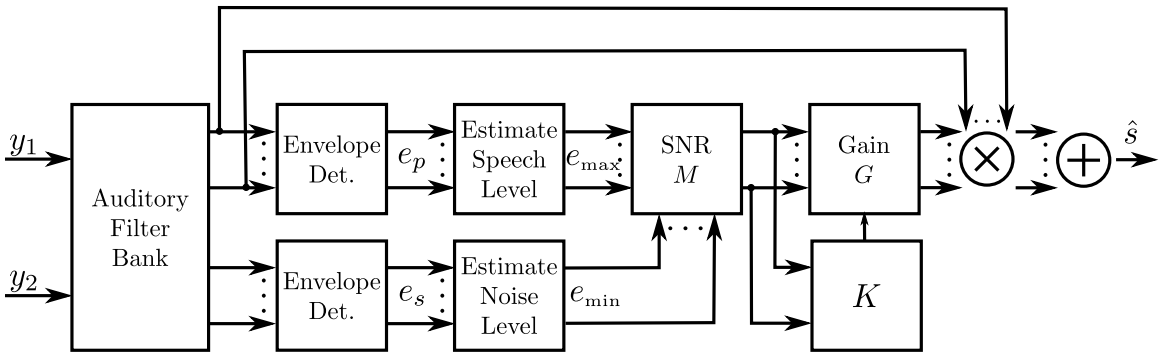


Figure 23: Block diagram of perceptual post processing where K is determined based on the *a priori* SNR. The time index has been dropped for brevity.

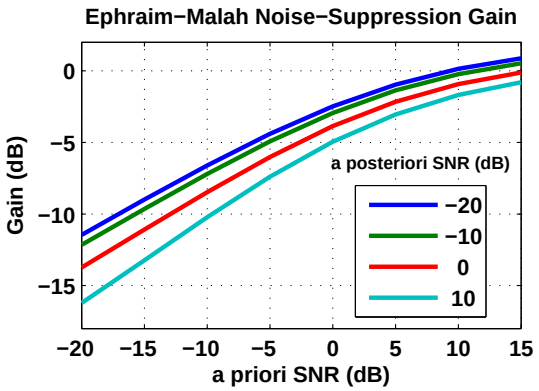
SNR, the signal is highly attenuated [12]. This behavior of the EMSR gain inherently reduces how fast the gain can change over time. Moreover, the *a priori* SNR is estimated using the decision-directed approach. This approach smooths the *a priori* SNR estimate at low SNR and, at high SNR, the *a priori* SNR follows the *a posteriori* SNR with a delay of one frame [12]. This smoothing of the *a priori* SNR also helps reduce the musical noise artifacts.

Using the same motivation of operating on different gain slopes depending on the mismatch between the *a priori* SNR and the *a posteriori* SNR, we propose a method to determine the value of K depending on the *a priori* SNR for the automatic gain control (AGC) noise suppression technique. A block diagram of this system is shown in Figure 23. The *a priori* SNR is estimated every 1 ms ms as

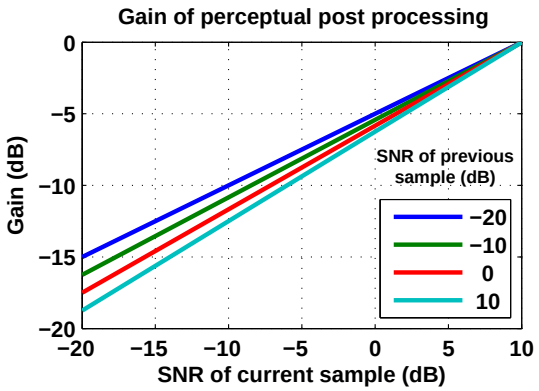
$$\begin{aligned}\mathcal{R}_{\text{prio}}[n] =& \alpha_{\text{speech}} \left(\frac{G_k[n-1] e_{p_k}[n]}{e_{k_{\min}}} \right)^2 \\ &+ (1 - \alpha_{\text{speech}}) \max \left(\left(\frac{e_{p_k}[n]}{e_{\min}} \right)^2, 0 \right)\end{aligned} \tag{46}$$

A maximum attenuation $K_{\max}$ is set such that it determines the gain slope for a set maximum *a priori* SNR $\max(\mathcal{R}_{\text{prio}})$. Similarly, a minimum attenuation $K_{\min}$ is set such that it determines the gain slope for a minimum *a priori* SNR $\min(\mathcal{R}_{\text{prio}})$. For our experiments, $\max(\mathcal{R}_{\text{prio}})$ is set to 10 dB, the $\min(\mathcal{R}_{\text{prio}})$ to -40 dB and the corresponding $K_{\min}$ to -15 dB and $K_{\max}$ to -20 dB. The K at which the gain operates on at any given time is given by

$$K = a\mathcal{R}_{\text{prio}} + b, \tag{47}$$



(a) EMSR gain vs. *a priori* SNR for different values of the *a posteriori* SNR



(b) AGC noise-suppression gain vs. *a priori* SNR for different values of the *a posteriori* SNR

Figure 24: Noise-suppression gain as a function of the SNR

where

$$a = \frac{K_{\min} - K_{\max}}{\min(\mathcal{R}_{\text{prio}}) - \max(\mathcal{R}_{\text{prio}})} \text{and}, \tag{48}$$

$$b = \frac{\min(\mathcal{R}_{\text{prio}})K_{\max} - \max(\mathcal{R}_{\text{prio}})K_{\min}}{\min(\mathcal{R}_{\text{prio}}) - \max(\mathcal{R}_{\text{prio}})}. \tag{49}$$

The variation of the gain as a function of the current SNR of the signal is shown in Figure 24(b) for different *a priori* SNR.

The limits of the *a priori* SNR $\mathcal{R}_{\text{prio}}$ does not seem to have an effect on the amount of musical noise perceived. These values of $\max(\mathcal{R}_{\text{prio}})$ and $\min(\mathcal{R}_{\text{prio}})$ were selected to cover the range of SNR observed in our experiments. However, if $K_{\min}$ and $K_{\max}$ are reduced to obtain more noise suppression, the amount of musical noise increases. This may be the case because along with the rate of change, in other words the modulation frequency, of the gain, the amount of modulation depth of the gain may also be a factor in the perception of musical noise.

5.4 Results

To quantify the the rate of change of the gain we calculate the Euclidean norm of $\frac{dG}{dt}$, which is given by

$$\|\nabla G\| = \sqrt{\left(\frac{dG}{dt}\right)^2}. \tag{50}$$

We can compare the rate of change of the gain for the cases where K is fixed over time and where K is given by (47) by comparing the values of $\mathcal{A} = \int_t \|\nabla G\| dt$ for both methods.

Figure 25 compares the value for $\mathcal{A}_{\text{var}}$ and $\mathcal{A}_{\text{fix}}$ for each subband of a speech sample that was corrupted by pub noise at 5 dB SNR. Figure 25(b) shows the values of $\mathcal{A}_{\text{var}}$ and $\mathcal{A}_{\text{fix}}$ for the noise-only periods. For all the frequency bands the value of $\mathcal{A}_{\text{var}}$ is lower than $\mathcal{A}_{\text{fix}}$. In Figure 25(c), we see these values of $\mathcal{A}_{\text{var}}$ and $\mathcal{A}_{\text{fix}}$ for the speech segments of the speech. For the critical band numbers 12—17 that corresponds to center frequencies between 613 Hz and 1924 Hz, where the speech is dominant, we see $\mathcal{A}_{\text{var}} > \mathcal{A}_{\text{fix}}$. This is supported by the fact that we observe some musical noise during the speech segments in the frequency bands in which speech is dominant. However, since the quality of speech is preserved because of the perceptual-based processing, the musical noise during these segments are not annoying.

We conducted a subjective test to determine if the proposed variable K technique reduces the perceived musical noise. Eight native English speakers were recruited. The subjects were asked to rate the sample based on the quality of speech, the amount of residual noise, and the amount of musical noise present on a scale of 1-to-5. They were presented with an example of clean, noisy, and speech corrupted with musical noise before the test. A total of 30 samples were presented to them. These samples included 24 samples processed by the BSS algorithm followed by the causal implementation of the perceptual post processing, 3 clean speech samples (Clean) and 3 mixtures (Mix) obtained from the microphone. Out of the 24 processed samples, 12 samples were processed using a fixed value of $K = 0.1$, which corresponds to $K = -20$ dB (Fix-K). The remaining 12 samples were processed by the proposed variable K given by (47) (Var-K). The values of $K_{\max}$ and $K_{\min}$ as described in Section 5.3.

	Speech Quality	Noise Level	Musical Noise
Var-K	3.41	3.36	3.06
Fix-K	3.28	3.01	2.32
Mix	3.33	1.04	4.16
Clean	4.95	5	4.95

Table 8: Results of the subjective test showing the average rating of each of the mentioned speech samples. The ratings are on a scale of 1-to-5, 1 being the worst and 5 being the best.

From Table 8, we can see that the proposed variable-K causal implementation of the

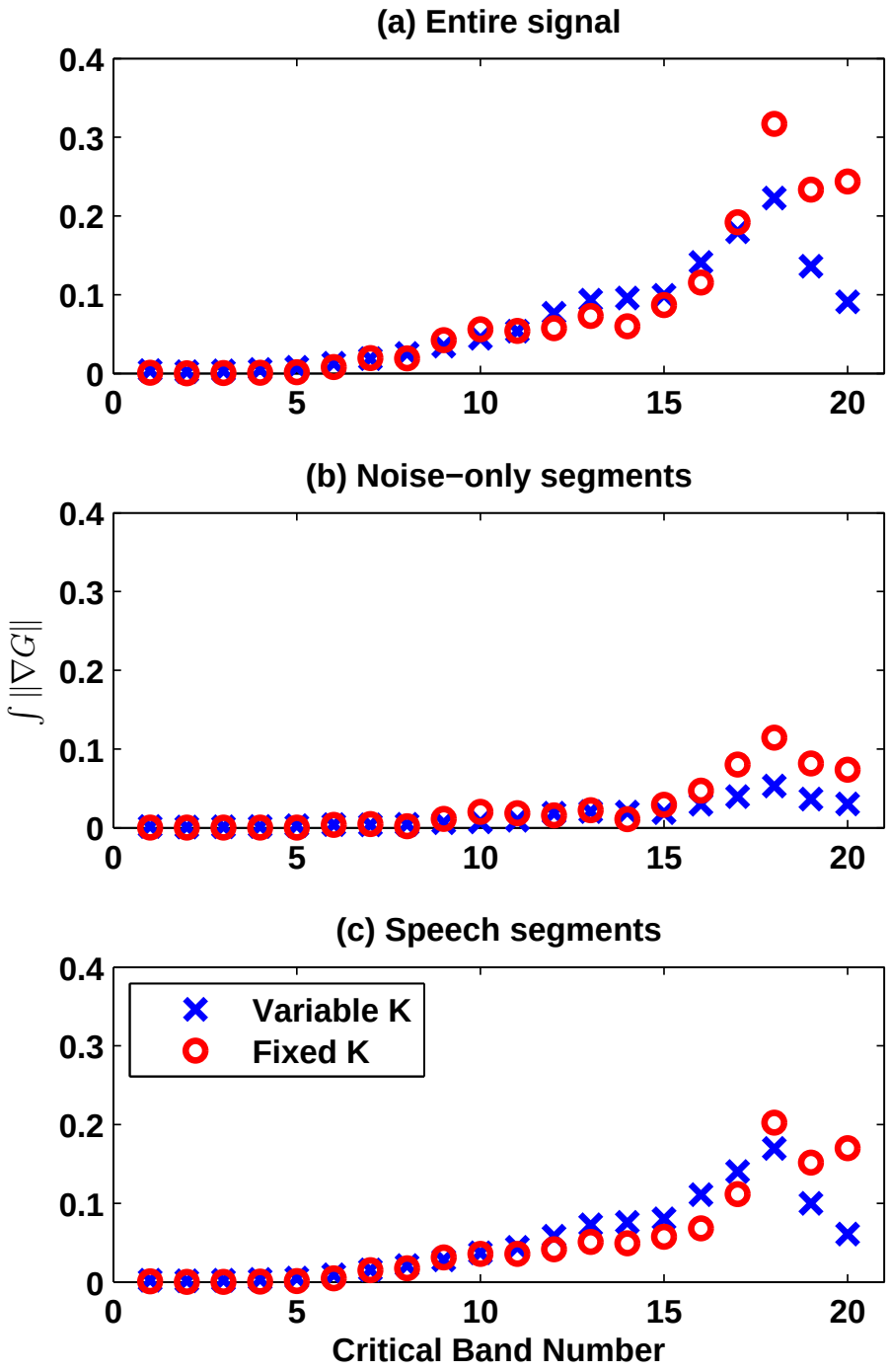


Figure 25: Area under the curves of $\|\nabla G\|$ for each subband of a speech sample that was corrupted by pub noise at 5 dB SNR.

perceptual post processing reduces the amount of musical noise.

In this chapter, we have presented a realistic implementation of a non-causal noise-suppression algorithm. In the process, we have addressed the causes of potential artifacts and shown that the rate of change of the gain is an important parameter that determines the amount of perceived musical noise. By reducing the rate of change of gain, we can reduce the amount of musical noise perceived. The proposed method of determining the expansion factor K based on the *a priori* SNR estimate implicitly smooths the gain. This implicit gain smoothing reduces the perceived musical noise, which is confirmed by the subjective tests.

CHAPTER VI

PERCEPTUAL OPTIMAL ESTIMATION

Traditional noise-suppression algorithms that are based on mathematical optimal estimation techniques may result in audible artifacts and speech distortion in the processed speech. The speech signal is analyzed into its spectral components using the Fourier transform. The noise-suppression gain is then obtained by minimizing the mean squared error between the estimate of the spectral components and the clean speech spectral components. Since the error that is minimized is not typically based on any perceptual quantity, the resulting noise suppressed speech may contain artifacts and speech distortions. The noise suppression gain is typically smoothed over time and frequency during noise-only segments to reduce the audibility of these artifacts. Hence, in the process to reduce the audible artifacts, we move away from the optimal estimate and use a solution that may not be mathematically optimal but sounds natural to the ear.

For most applications for which the noise-suppressed speech is meant to be heard by humans, it is not necessary to exactly reconstruct the clean-speech signal in a statistical sense. It should be sufficient to suppress the noise present in the noisy speech such that the noise-suppressed speech signal approximates the desired clean-speech signal in a perceptual sense. In this chapter, first, we use the envelopes of the critical bands to estimate the spectral components of the signal. These estimates are then used in the state-of-the-art noise-suppression gains, Ephraim-Malah suppression rule (EMSR) and the Wiener gain. We show that this type of processing reduces the perceptibility of artifacts. Later, we derive both non-linear and linear optimal estimators that operate in the perceptual domain. These gains are such that the resulting speech sounds natural without any further tweaking of the gain.

6.1 Using Envelopes as Estimates of the Spectral Components

In this section, we combine perceptual-based processing with standard noise suppression techniques. Instead of estimating the spectral components of the signal using a FFT, we use the envelopes of the critical bands to get an idea of the spectral content of the signal. The two standard noise suppression techniques that we use here are: Wiener gain, and Ephraim-Malah suppression rule.

6.1.1 Analysis of Noisy Speech

Let $x(t)$ denote the noisy speech signal which can be written as,

$$x(t) = s(t) + n(t), \tag{51}$$

where, $s(t)$ is the clean speech signal that has been corrupted by noise $n(t)$. The signal is analyzed using a cochlea filter bank and the envelope extraction technique described in Section 3.2. Moreover, the square of the envelope $e_i^2(t)$ of the signal indicates how the energy of the channel changes over time. This can be used to estimate the spectral content of the signal. The *a priori* SNR [18] of the i-th channel of the noisy signal can be given by,

$$\xi_i = \frac{e_{s_i}^2(t)}{\sigma_n^2}, \tag{52}$$

where, $e_{s_i}(t)$ is the envelope of the i-th channel of the clean speech $s(t)$ and σ_n^2 the energy of of the noise. The *a posteriori* SNR [18] of the i-th channel of the noisy signal is given by,

$$\gamma_i = \frac{e_{x_i}^2(t)}{\sigma_{n_i}^2}, \tag{53}$$

where, $e_{x_i}(t)$ is the envelope of the i-th channel of the noisy speech $x(t)$. The *a priori* and *a posteriori* SNR can also be calculated on a frame-by-frame basis. In this case the SNR will be denoted as $\xi_{i,k}$ and $\gamma_{i,k}$, where k is the index of the time frame.

The *a priori* SNR ξ_i of the signal is the true SNR of the signal. Given only the noisy speech signal $x(t)$, this SNR can only be estimated since neither the energy of the clean speech nor the energy of the noise is available. The *a posteriori* SNR can be calculated exactly only if the noise energy is available.

6.1.2 Noise Estimation

In the AGC-based noise-suppression technique described in Chapter 3, if the noise energy is underestimated (i.e. the actual noise energy is more than the estimate), the gain may not be as effective but it will not distort the speech. However, in the Wiener gain based noise suppression and the Ephraim-Malah suppression rule, the gain is calculated on the assumption that the noise energy is known. These noise suppression gains are more sensitive to errors in noise estimation. If the noise is estimated incorrectly then the final speech may be heavily distorted. In Chapter 3, we use the minimum of the envelope to estimate the noise floor. In contrast here, we process each channel of the noisy signal on a frame-by-frame basis. A voice activity detector (VAD) is used to determine, in a time frame, whether speech is present. The noise energy can be calculated as the average energy of the frame in which there is no speech. This type of calculation gives a more accurate estimate of the noise energy present in the noisy signal.

$$\hat{\sigma}^2_{n_i} = \mu\hat{\sigma}^2_{n_i} + (1-\mu)\text{mean}(e^2_{x_{i,k}}(t)) \tag{54}$$

where, k is the index of the time frame when there is no speech and μ is a smoothing factor such that $\mu < 1$. The smoothing factor prevents the estimate of the energy from changing too drastically between time frames, but at the same time allows the estimate to adapt slowly to changing levels of noise. For our experiments μ is set to 0.8.

6.1.3 SNR Estimation

The noise estimate is calculated using (54) and then the estimate of the *a posteriori* SNR $\hat{\gamma}_{i,k}$ can be calculated using (53) as,

$$\hat{\gamma}_{i,k} = \frac{e^2_{x_{i,k}}(t)}{\hat{\sigma}^2_{n_i}} \tag{55}$$

The *a priori* SNR can then be estimated as,

$$\hat{\xi}_{i,k} = \alpha\frac{G_{i,k-1}e^2_{x_{i,k-1}}(t)}{\hat{\sigma}^2_{n_i}} + (1-\alpha)P[\hat{\gamma}_{i,k} - 1] \tag{56}$$

where, $P[x] = x$ if $x \geq 0$ and $P[x] = 0$ otherwise. α is a smoothing factor [18, 12].

In [18], 256 samples are processed at a time at 8 kHz sampling frequency. Each frame overlaps the previous frame by 192 samples. Hence, at each frame only 8 ms of new data contribute to the calculation of the SNR. The smoothing factor α is set to 0.98. At this value of α it takes 34 such 8 ms hops for the SNR to reach 50% of its original value. In other words, the *a priori* SNR estimate is smoothed over 0.25 s. In our processing, we analyze 20 ms of data at a time with 50% overlap at 16 kHz. To obtain the same amount of smoothing of the *a priori* SNR over time, we set α to 0.9727.

6.1.4 Gain Computation

Wiener Gain for Noise Suppression The noise energy is estimated using (54). The gain G_{Wiener} [51] for the i-th channel and the k-th time frame can be calculated as,

$$G_{\text{Wiener}_{i,k}} = \sqrt{\frac{\hat{\xi}_{i,k}}{1+\hat{\xi}_{i,k}}}, \tag{57}$$

where, $\hat{\xi}_{i,k}$ is the estimate of the *a priori* SNR calculated using (56).

Ephraim-Malah Suppression Rule (EMSR) The noise energy, *a posteriori* and *a priori* SNR is calculated using (54), (55) and (56) respectively. The EMSR gain G_{EMSR} [18] for the i-th channel and the k-th time frame is calculated as,

$$\begin{aligned} G_{\text{EMSR}_{i,k}} = \frac{\sqrt{\pi}}{2} \sqrt{\frac{1}{1+\hat{\gamma}_{i,k}} \cdot \frac{\hat{\xi}_{i,k}}{1+\hat{\xi}_{i,k}}} \\ \times \text{M}\left[(1+\hat{\gamma}_{i,k}) \cdot \frac{\hat{\xi}_{i,k}}{1+\hat{\xi}_{i,k}}\right] \end{aligned} \tag{58}$$

where, M[$\cdot$] is defined as,

$$\text{M}[\theta] = \exp\left(-\frac{\theta}{2}\right)\left[(1+\theta)I_0\left(\frac{\theta}{2}\right) + \theta I_1\left(\frac{\theta}{2}\right)\right] \tag{59}$$

where I_0 and I_1 are the modified Bessel functions of the zeroth and first order.

6.1.5 Results

The spectrograms of the unprocessed and processed speech samples are shown in Fig. 26. The speech processed using FFT analysis distorts the speech and the background noise. The level of noise in the processed speech is lowered but at the same time the resulting

speech sounds unnatural and the residual noise is annoying to the ear. However, the speech processed using the critical band analysis uniformly decreases the background noise without distorting the speech.

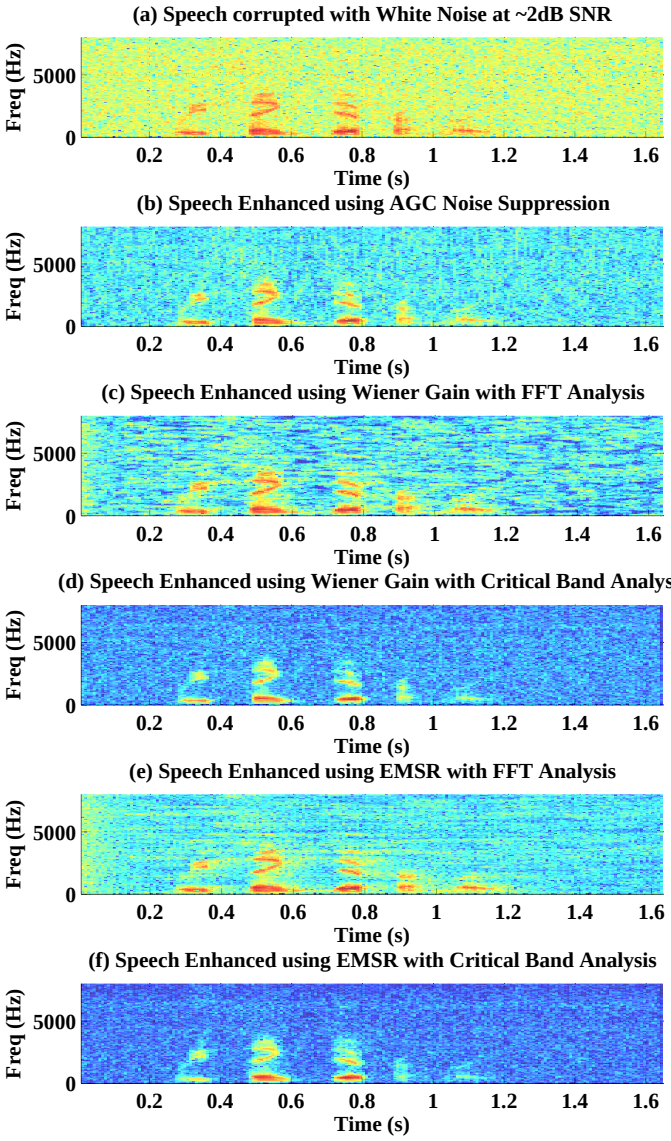


Figure 26: Spectrogram of (a) Speech corrupted with white noise at -2dB SNR, (b) noise suppressed speech using AGC noise suppression, (c) noise suppressed speech using Wiener gain with FFT analysis, (d) noise suppressed speech using Wiener gain with critical band analysis, (e) noise suppressed speech using EMSR with FFT analysis, (d) noise suppressed speech using EMSR with critical band analysis

A listening test was conducted to evaluate the performance of the proposed signal analysis. Twelve native English speaking subjects were recruited. Two samples were presented at each trial. The speech samples were corrupted using babble noise, white noise and pink noise at 2dB, 5dB and 10dB SNR. Noise suppressed samples were presented to the subjects in pairs. These samples were processed using the AGC-based noise suppression, Wiener Gain and the EMSR. The sound samples processed using the Wiener gain and EMSR with critical band analysis was compared to the corresponding FFT based processing. The AGC-based noise suppression was compared to the Wiener gain and EMSR gain using critical bands. Nine samples were presented to the subjects for each type of comparison which results to a total of thirty-six samples. The subjects were asked to rate the second sample compared to the first on the basis of quality of speech and noise level. They were asked to rate the quality of speech depending on how natural the speech sounded and the amount of distortions present in the speech. The possible ratings for quality of speech were that the second sample was much better (3), better (2), slightly better (1), about the same (0), slightly worse (-1), worse (-2) and much worse (-3) than the first sample. They were asked to rate the overall noise level in terms of how annoying was the residual noise. The allowable responses in this category were the residual noise in the second sample was less annoying (2), slightly less annoying (1), about the same (0), slightly more annoying (-1) and more annoying (-2) than the first sample. The samples were presented to the subjects in a random order and no information about the nature of the samples were given to the subjects.

	Speech Quality	Noise Level
EMSR FFT	0.7	1.4

Table 9: Results of the subjective test showing the average rating of the performance of EMSR using critical bands analysis compared to EMSR using FFT analysis. The ratings are on a scale of -3-to-3, -3 corresponds to much worse and 3 corresponds to much better.

The results of the subjective test are shown in Table 9 - 11. The Wiener gain and EMSR using critical band analysis result in a residual noise that is less annoying and more

	Speech Quality	Noise Level
Wiener FFT	1	1.3

Table 10: Results of the subjective test showing the average rating of the performance of Wiener gain using critical bands analysis compared to EMSR using FFT analysis. The ratings are on a scale of -3-to-3, -3 corresponds to much worse and 3 corresponds to much better.

	Speech Quality	Noise Level
Wiener Critical Band	0.75	-0.9
EMSR Critical Band	0.6	-1.3

Table 11: Results of the subjective test showing the average rating of the performance of AGC-based noise suppression compared to Wiener gain and EMSR using critical band analysis. The ratings are on a scale of -3-to-3, -3 corresponds to much worse and 3 corresponds to much better.

natural to listen to. The AGC-based noise suppression is not as aggressive as the Wiener-gain-based noise suppression and EMSR noise suppression but the resulting speech has a better quality. In Chapter 7, we discuss in more detail the cause of the musical noise in the perceptual-based processing.

6.2 Non-linear Minimum Mean-square Estimators

The Wiener and the EMSR gain are optimal FFT spectral estimators. Using the critical band analysis and the envelopes as estimates of the spectral content with the Wiener and EMSR gain is a rough fit and may not be optimal. We, now, derive an optimal estimator to estimate the clean speech envelope from the noisy speech envelope.

Let us assume the envelopes in each subband are independent of each other and

$$e_{x_i} = e_{s_i} + e_{n_i}, \tag{60}$$

where e_{x_i}, e_{s_i}, and e_{n_i} are the envelopes of the noisy speech, clean speech, and noise in the i-th frequency band respectively. To simplify the notations, we drop the subscript i to indicate the envelope of a particular frequency band. We can estimate the envelope of the clean speech by minimizing the following error

$$\xi = \mathbf{E}\{(e_s - \hat{e}_s)^2\}, \tag{61}$$

where e_s is the true clean speech envelope, $\hat{e}_s$ is the estimated clean speech envelope, and ξ is the mean-squared error in the i-th frequency band.

The non-linear estimator that minimizes ξ is given by

$$\begin{aligned}\hat{e}_s &= \mathbf{E}\{e_s|e_x\} \\ &= \int_0^\infty e_s f(e_s|e_x)\, de_s,\end{aligned} \tag{62}$$

where e_x is the noisy speech envelope in i-th frequency band, and $f(\cdot)$ is the probability density function (PDF) of its argument. Here, the estimated clean envelope $\hat{e}_s$ in subband i is given by the condition expectation of the clean envelope e_s given the noisy envelope e_x. Now, the conditional PDF $f(e_s|e_x)$ can be written as

$$\begin{aligned}f(e_s|e_x) &= \frac{f(e_s, e_x)}{f(e_x)}, \\ &= \frac{f(e_x|e_s)f(e_s)}{f(e_x)},\end{aligned} \tag{63}$$

where

$$f(e_x) = \int_0^\infty f(e_x|e_s)f(e_s)\, de_s. \tag{64}$$

Now, the conditional probability $f(e_x|e_s)$ can be written as

$$f(e_x|e_s) = \frac{f(e_s, e_x)}{f(e_s)}. \tag{65}$$

The joint probability $f(e_s, e_x)$, by the theory of two functions of two random variables [38], can be written as

$$f(e_s, e_x) = f(e_s) \cdot f_{e_n}(e_x - e_s), \tag{66}$$

where $f_{e_n}(\cdot)$ is the PDF of the noise envelope.

By substituting equations (63)—(66) in (62), we get

$$\hat{e}_s = \frac{\int_0^\infty e_s f_{e_n}(e_x - e_s) \cdot f(e_s)\, de_s}{\int_0^\infty f_{e_n}(e_x - e_s) \cdot f(e_s)\, de_s}. \tag{67}$$

Hence, an estimate of the clean speech can be obtained by modeling the PDF of the speech and noise envelopes. In the next sections, we model the envelope of the signal as Gaussian, and as gamma random variables, and simplify (67) to obtain the estimated clean speech envelope $\hat{e}_s$.

6.2.1 Gaussian-distributed Envelopes

Assume that the envelopes of the speech subbands are half-normal distributed

$$f(e_s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left(-\frac{e_s^2}{2\sigma_s^2}\right), \tag{68}$$

where $\mathrm{E}(e_s) = \sigma_s\sqrt{\frac{2}{\pi}}$ and $\mathrm{Var}(e_s) = \sigma_s^2\left(1 - \frac{2}{\pi}\right)$. Also, assume that the envelopes of the noise subbands are one-sided Gaussian distributed

$$f(e_n) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(\frac{-(e_n - \mu_n)^2}{2\sigma_n^2}\right) \tag{69}$$

where $\mathrm{E}(e_n) = \mu_n$ and $\mathrm{Var}(e_n) = \sigma_n^2$.

Subtituting (65), (66), (68), and (69) in (64), we get

$$\begin{aligned} f(e_x) &= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{((e_x - e_s) - \mu_n)^2}{2\sigma_n^2}\right) \cdot \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left(-\frac{e_s^2}{2\sigma_s^2}\right) de_s \\ &= \frac{1}{2\pi\sigma_n\sigma_s} \int_0^\infty \exp\left(-\frac{((e_x - \mu_n) - e_s)^2}{2\sigma_n^2}\right) \cdot \exp\left(-\frac{e_s^2}{2\sigma_s^2}\right) de_s. \end{aligned} \tag{70}$$

Let $\tilde{e}_s = e_x - \mu_n$. Therefore, (70) can be written as

$$\begin{aligned} f(e_x) &= \frac{1}{2\pi\sigma_n\sigma_s} \int_0^\infty \exp\left(-\frac{(\tilde{e}_s - e_s)^2}{2\sigma_n^2}\right) \cdot \exp\left(-\frac{e_s^2}{2\sigma_s^2}\right) de_s \\ &= A_1 \int_0^\infty \exp\left(-a_1 e_s^2 + b_1 e_s\right) de_s, \end{aligned} \tag{71}$$

where

$$\begin{aligned} A_1 &= \frac{1}{2\pi\sigma_n\sigma_s} \exp\left(-\frac{\tilde{e}_s^2}{2\sigma_n^2}\right), \\ a_1 &= \left(\frac{1}{2\sigma_n^2} + \frac{1}{2\sigma_s^2}\right), \\ b_1 &= \left(\frac{\tilde{e}_s}{\sigma_n^2}\right). \end{aligned} \tag{72}$$

We can further simplify (71) as follows

$$\begin{aligned} f(e_x) &= A_1 \int_0^\infty \exp\left(-a_1 e_s^2 + b_1 e_s\right) de_s, \\ &= A_1 \exp\left(\frac{b_1^2}{4a_1}\right) \int_0^\infty \exp\left(-a_1\left(e_s - \frac{b}{2a}\right)\right) de_s \\ &= A_1 \exp\left(\frac{b_1^2}{4a_1}\right) \frac{1}{2}\sqrt{\frac{\pi}{a_1}} \end{aligned} \tag{73}$$

Now, the numerator of (67) is

$$\text{Numerator} = \int_0^\infty e_s f_{e_n}(e_x - e_s) \cdot f(e_s)\, de_s, \tag{74}$$

which can be simplified to

$$\begin{aligned}
\text{Numerator} &= A_1 \int_0^\infty e_s \exp\left(-a_1 e_s^2 + b_1 e_s\right)\, de_s \\
&= A_1 \exp\left(\frac{b_1^2}{4a_1}\right) \int_0^\infty e_s \exp\left(-a_1\left(e_s - \frac{b_1}{2a_1}\right)\right)\, de_s \\
&= \exp\left(\frac{b_1^2}{4a_1}\right)\left[\frac{1}{2a_1} + \frac{b_1}{4a_1}\sqrt{\frac{\pi}{a_1}}\right].
\end{aligned} \tag{75}$$

Substituting (75) and (73) in (67)

$$\begin{aligned}
\hat{e}_s &= \frac{\frac{1}{2a_1} + \frac{b_1}{4a_1}\sqrt{\frac{\pi}{a_1}}}{\frac{1}{2}\sqrt{\frac{\pi}{a_1}}} \\
&= \frac{1}{\sqrt{\pi a_1}} + \frac{b_1}{2a_1}
\end{aligned} \tag{76}$$

From (72), (76) becomes

$$\begin{aligned}
\hat{e}_s &= \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma_n \sigma_s}{\sqrt{\sigma_n^2 + \sigma_s^2}} + \frac{\tilde{e}_s \sigma_s^2}{\sigma_n^2 + \sigma_s^2} \\
&= \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma_s}{\sqrt{1 + \frac{\sigma_s^2}{\sigma_n^2}}} + (e_x - \mu_n) \cdot \frac{\frac{\sigma_s^2}{\sigma_n^2}}{1 + \frac{\sigma_s^2}{\sigma_n^2}},
\end{aligned} \tag{77}$$

where $\frac{\sigma_s^2}{\sigma_n^2}$ is the SNR of the noisy signal, and $(e_x - \mu_n)$ is the spectral-subtraction estimate of the clean speech envelope.

Modeling the envelopes as Gaussian random variables may not be an accurate statistical model, but results in an estimate of the clean speech envelope that is easy to compute.

6.2.2 Gamma-distributed Envelopes

Assume that the envelopes of the speech subbands are Gamma distributed

$$f(e_s) = \frac{e_s^{(\alpha-1)}}{\theta^\alpha \Gamma(\alpha)} \exp\left(\frac{-e_s}{\theta}\right), \tag{78}$$

where $\mathrm{E}(e_s) = \alpha\theta$ and $\mathrm{Var}(e_s) = \alpha\theta^2$. Here, α is called the shape parameter and θ is called the scale parameter. Also, assume that the envelopes of the noise subbands are Gaussian

distributed

$$f(e_n) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(e_n-\mu)^2}{2\sigma^2}\right) \tag{79}$$

where $\mathrm{E}(e_n) = \mu$ and $\mathrm{Var}(e_n) = \sigma^2$.

Subtituting (65), (66), (78), and (79) in (64), we get

$$\begin{aligned} f(e_x) &= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-((e_x-e_s)-\mu)^2}{2\sigma^2}\right) \cdot \frac{e_s^{(\alpha-1)}}{\theta^\alpha \Gamma(\alpha)} \exp\left(\frac{-e_s}{\theta}\right) de_s \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \frac{1}{\theta^\alpha\Gamma(\alpha)} \cdot \exp\left(\frac{-(e_x-\mu)^2}{2\sigma^2}\right) \int_0^\infty e_s^{\alpha-1} \exp\left(-\frac{e_s^2}{2\sigma^2} - \left(\frac{1}{\theta} - \frac{(e_x-\mu)}{\sigma^2}\right) e_s\right) de_s \\ &= A_2 \int_0^\infty e_s^{\alpha-1} \exp\left(-\frac{e_s^2}{2\sigma^2} - \left(\frac{1}{\theta} - \frac{(e_x-\mu)}{\sigma^2}\right) e_s\right) de_s \end{aligned} \tag{80}$$

where

$$A_2 = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \frac{1}{\theta^\alpha\Gamma(\alpha)} \cdot \exp\left(\frac{-(e_x-\mu)^2}{2\sigma^2}\right). \tag{81}$$

Now,

$$\int_0^\infty x^{p-1} \exp\left(-\beta x^2 - \gamma x\right) dx = (2\beta)^{(-p/2)} \Gamma(p) \exp\left(\frac{\gamma^2}{8\beta}\right) D_{-p}\left(\frac{\gamma}{\sqrt{2\beta}}\right), \tag{82}$$

where $D_p(z)$ is a parabolic cylinder function, and $\Gamma(p)$ is a gamma function [60].

Hence, (80) can be written as

$$f(e_x) = A_2 \left(\frac{1}{\sigma^2}\right)^{\left(\frac{-\alpha}{2}\right)} \Gamma(\alpha) D_{-\alpha}\left(\left(\frac{1}{\theta} - \frac{(e_x-\mu)}{\sigma^2}\right)\sigma\right) \tag{83}$$

Now, the numerator of (67) is

$$\text{Numerator} = \int_0^\infty e_s \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-((e_x-e_s)-\mu)^2}{2\sigma^2}\right) \cdot \frac{e_s^{(\alpha-1)}}{\theta^\alpha\Gamma(\alpha)} \exp\left(\frac{-e_s}{\theta}\right) de_s, \tag{84}$$

which can be similarly simplified to

$$\begin{aligned} \text{Numerator} &= A_2 \int_0^\infty e_s^\alpha \exp\left(-\frac{e_s^2}{2\sigma^2} - \left(\frac{1}{\theta} - \frac{(e_x-\mu)}{\sigma^2}\right) e_s\right) de_s \\ &= A_2 \left(\frac{1}{\sigma^2}\right)^{\left(-\frac{\alpha+1}{2}\right)} \Gamma(\alpha+1) D_{-\alpha}\left(\left(\frac{1}{\theta} - \frac{(e_x-\mu)}{\sigma^2}\right)\sigma\right) \end{aligned} \tag{85}$$

Therefore, (67) becomes

$$\hat{e}_s = \frac{\left(\frac{1}{\sigma^2}\right)^{\left(-\frac{\alpha+1}{2}\right)} \Gamma(\alpha+1) D_{-(\alpha+1)}\left(\left(\frac{1}{\theta} - \frac{(e_x-\mu)}{\sigma^2}\right)\sigma\right)}{\left(\frac{1}{\sigma^2}\right)^{\left(-\frac{\alpha}{2}\right)} \Gamma(\alpha) D_{-\alpha}\left(\left(\frac{1}{\theta} - \frac{(e_x-\mu)}{\sigma^2}\right)\sigma\right)} \tag{86}$$

Now,

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$$

$$D_{p+1}(z) = zD_p(z) + pD_{p-1}(z)$$

$$\therefore D_{-\alpha+1}(z) = zD_{-\alpha}(z) - \alpha D_{-(\alpha+1)}(z)$$

$$\text{and let} \quad z = \left(\frac{1}{\theta} - \frac{(e_x-\mu)}{\sigma^2}\right)\sigma$$

Hence, (86) becomes

$$\begin{aligned}\hat{e}_s &= \left(\frac{1}{\sigma^2}\right)^{-1/2} \cdot \alpha \left[\frac{\frac{1}{\alpha}\left[zD_{-\alpha}(z) - D_{-\alpha+1}(z)\right]}{D_{-\alpha}(z)}\right] \\ &= \left(\frac{1}{\sigma^2}\right)^{-1/2} \left[z - \frac{D_{-\alpha+1}(z)}{D_{-\alpha}(z)}\right] \\ &= \sigma \cdot \left[z - \frac{D_{-\alpha+1}(z)}{D_{-\alpha}(z)}\right] \qquad (87)\end{aligned}$$

(88)

Let $\tilde{e}_s = e_x - \mu$, then (88) becomes

$$\hat{e}_s = \frac{\sigma^2}{\theta} - \tilde{e}_s + \sigma \cdot \frac{D_{-\alpha+1}\left(\frac{\sigma}{\theta} - \frac{\tilde{e}_s}{\sigma}\right)}{D_{-\alpha}\left(\frac{\sigma}{\theta} - \frac{\tilde{e}_s}{\sigma}\right)} \qquad (89)$$

6.3 Linear Minimum Mean-square Estimators

As seen in the previous section, non-linear MMSE estimators can be computationally complex to implement. However, if we constrain the estimator to be linear, we may be able to obtain an estimator that is relatively simple to compute. Moreover, a linear estimator will prevent the noise-suppression gain from becoming unstable. In this section, we derive a linear-MMSE estimator that estimates the clean speech signal by minimizing the error between the estimated loudness and the true loudness of the signal in the perceptual domain.

We can estimate the loudness of the signal by taking the logarithm of the envelope of the critical bands. While this transform is not an exact representation of the loudness of the signal, it is sufficient to obtain an approximate estimate. Moreover, a linear gain in the log-domain non-linearly expands the dynamic range of the envelope. This expansion ensures the

lowering the background noise in a perceptual sense rather than of the completely removing the background noise. Let,

$$\log \hat{e}_s = A \log e_x + \log b, \tag{90}$$

where $\hat{e}_s$ is the estimate of the clean-speech envelope, e_x is the noisy envelope, and A and $\log b$ are the the gain parameters. The subscript i indicating the subband of operation is dropped for convenience. Hence,

$$\hat{e}_s = b e_x^A. \tag{91}$$

This equation is of the same form as (13). As seen previously, the gain that is applied to the subband channel can be written as

$$G = b e_x^{(A-1)}. \tag{92}$$

In contrast to Chapter 3, in this section we obtain the gain parameters by minimizing the following error

$$\xi = \mathbf{E}\{(\log \hat{e}_s - \log e_x)^2\}. \tag{93}$$

From (90), the mean-squared error (MSE) can be written as

$$\xi = \mathbf{E}\{(A \log e_x + \log b - \log e_x)^2\}. \tag{94}$$

The gain parameters A and $\log b$ that minimize ξ can be obtained by differentiating (94) with respect to A and $\log b$ and equating this differentiation to zero. Hence,

$$\begin{aligned}
\frac{\partial \xi}{\partial A} &= 0 \\
\mathbf{E}\left\{2\left(A \log e_x + \log b - \log e_s\right) \log e_x s\right\} &= 0 \\
2A\mathbf{E}\{(\log e_x)^2\} + 2 \log b \mathbf{E}\{\log e_x\} - 2\mathbf{E}\{\log e_s \log e_x\} &= 0 \\
A\mathbf{E}\{(\log e_x)^2\} + \log b \mathbf{E}\{\log e_x\} - \mathbf{E}\{\log e_s \log e_x\} &= 0
\end{aligned} \tag{95}$$

$$\begin{aligned}
\frac{\partial \xi}{\partial \log b} &= 0 \\
\mathbf{E}\left\{2\left(A \log e_x + \log b - \log e_s\right)\right\} &= 0 \\
2A\mathbf{E}\{\log e_x\} + 2 \log b - 2\mathbf{E}\{\log e_s\} &= 0 \\
A\mathbf{E}\{\log e_x\} + \log b - \mathbf{E}\{\log e_s\} &= 0
\end{aligned} \tag{96}$$

Let

$$\begin{aligned}
\mathbf{E}\{\log e_x\} &= m_x, \\
\mathbf{E}\{\log e_s\} &= m_s, \\
\mathbf{E}\{(\log e_x)^2\} &= c_x, \quad \text{and} \\
\mathbf{E}\{\log e_x \log e_s\} &= r_{xs}.
\end{aligned} \tag{97}$$

Therefore, (95) and (96) can be written as

$$\begin{aligned}
Ac_x + \log bm_x - r_{xs} &= 0 \\
Am_x + \log b - m_s &= 0
\end{aligned}$$

Solving, the above two equations we get

$$A = \frac{r_{xs} - m_x m_s}{c_x - m_x^2} \tag{98}$$

$$\log b = m_s - Am_x \tag{99}$$

6.3.1 Proof-of-concept Implementation

As a proof of concept, we first implement this linear-in-the-log-domain MMSE assuming that the statistics of the signal can be calculated exactly. We calculate m_s and r_{xs} for each subband assuming the clean envelope is available. The gain parameters are computed using (98) and (99).

The statistics of the signal were calculated on a frame-by-frame basis and the statistics were updated every 10 ms. The gain parameters were also updated every frame. In this case, the noise is entirely removed, but the higher frequency bands where speech is entirely masked by the noise, the gain shapes the noise, spectrally and temporally, to speech. Hence, the carrier of the speech in the higher frequency bands is narrow band noise instead of periodic pulses.

We did not purse this avenue further because a realistic implementation of such a system involves estimating the statistics of the signal, which is beyond the scope of this thesis.

CHAPTER VII

MUSICAL NOISE ANALYSIS

In this thesis, we have shown that we can operate in the critical bands and modify the dynamic range of the envelope in the subbands to suppress noise. Since this technique is similar to the processing done by the peripheral auditory system, the processing should be more perceptually transparent compared to other traditional noise-suppression methods. However, in our proposed methods, when the SNR of the noisy signal is very low (< 5 dB) the perceptual quality of the residual noise is not the same as the noise present in the noisy signal. The residual noise is modified and may sound like musical noise. The musical noise that is heard as a result of the proposed processing can be described as modulated tones perceived at different times and frequencies.

Musical noise is a common artifact of speech processing techniques. The spurious peaks that remain in the enhanced spectrum while suppressing the noise using spectral subtraction are heard as artifacts, which is also described as musical noise [9]. The circular convolution artifacts that arise from processing the signal in overlapping blocks is also heard as musical noise [35].

Depending on the cause of the musical noise, there are various techniques to reduce its effect. Spectral over subtraction is one technique used to reduce the musical noise in spectral subtraction type noise-suppression algorithms [9]. In this technique, the background noise is removed such that there is a residual noise floor, which masks the musical noise. Moreover, the noise-suppression gain can smoothed over time and frequency to reduce the musical noise artifacts. The decision-directed approach to estimate the *a priori* SNR that is used in the Ephraim-Malah suppression rule achieves such gain smoothing [12]. Filters can also be designed by morphologically processing the spectrogram to removed the time-frequency regions that contain musical noise [22]. Since musical noise can arise as a result of different types of processing and for each there is a different approach to reduce the musical noise,

it may not possible to apply existing methods to reduce the musical noise seen in the perceptual noise suppression.

Anderson, in [4], describes existing noise-suppression algorithms that try to reduce musical noise as *placing an ambulance at the bottom of a cliff* approach, where the damage has been done, and efforts are made to reduce the impact. Instead, we are interested in removing the musical noise by *building a fence at the top of the hill*. To achieve this goal, we need to first understand the cause of the musical noise in the perceptual-based processing. The combination of complex signals and different knobs that can be tuned makes it difficult to try all possible permutations to see which set of parameters does not generate musical noise. Moreover, the musical noise is not heard in each individual critical band, but is heard only when the critical bands are combined. This phenomenon makes it difficult to investigate the cause of musical noise in each band independently where the number of parameters are limited.

In this chapter, we will investigate the cause of this musical noise, and propose techniques that can reduce the perception of this musical noise.

7.1 Dynamic Range Expansion of Clean Speech and Noise

To show that the non-linear expansion of the dynamic range of clean speech is perceptually transparent, we analyze the clean speech using the setup described in Section 3.2. Hence, the clean speech $s[n]$ can be expressed by equation (8). Recall, the signal $s[n]$ is decomposed into critical bands $c_i[n]$ using a constant-Q filter bank, which is described in Section 3.2. The envelope $e_i[n]$ in each critical band is extracted using a non-linearity/low-pass filter (LPF) combination, which is described in Section 3.2. In this chapter, we will indicate this envelope as $e_{\text{LPF}}[n]$. In each critical band for the entire clean speech sample, the long-term maximum $e_{\max}$ and minimum of envelopes $e_{\min}$ are computed as

$$\begin{aligned} e_{\max} &= \max(e_{\text{LPF}}[n]) \\ e_{\min} &= \min(e_{\text{LPF}}[n]). \end{aligned} \tag{100}$$

As in equation (18), $M = \frac{e_{\max}}{e_{\min}}$ indicates the dynamic range of the signal.

Recall, the subscript i, which indicates the critical band number, has been dropped for convenience. The envelope of the clean speech is modified according to equation (13). Rewriting equation (13)

$$\hat{e}_{\mathrm{LPF}}[n] = \beta e^{\alpha}_{\mathrm{LPF}}[n], \tag{101}$$

where β and α are calculated according to equations (16) and (17), which are

$$\alpha = 1 - \frac{\log K}{\log M} \tag{102}$$

and

$$\beta = e^{1-\alpha}_{\max}, \tag{103}$$

where K is set to 0.01, and $M = \frac{e_{\max}}{e_{\min}}$. Since $K < 1$, the dynamic range of the envelope is expanded. This expansion can be written as a multiplicative gain, as per equation (14)

$$\hat{e}_{\mathrm{LPF}}[n] = G[n] e_{\mathrm{LPF}}[n], \tag{104}$$

where

$$G[n] = \beta e^{\alpha-1}_{\mathrm{LPF}}[n]. \tag{105}$$

The gain $G[n]$ that expands the dynamic range of the critical band is applied to the critical band $c[n]$.

Take the logarithm of equation (101), we get

$$\log \hat{e}_{\mathrm{LPF}}[n] = \alpha \log e_{\mathrm{LPF}}[n] + \log \beta. \tag{106}$$

Hence, the envelope modification in each critical band is an affine transformation in the log domain where the slope of the modification is α and the y-intercept is $\log \beta$. The slope α also indicates the expansion ratio of envelope mapping.

For this type of processing for clean speech the perceived quality of speech is almost unaltered. The spectrogram of the clean speech before and after the dynamic range expansion is shown in Figure 27.

However, when a noise-only signal is processed in a similar fashion, the output contains the musical noise that we hear in results presented in this thesis. In Figure 28, the spectrogram of white noise before and after the dynamic range expansion is shown.

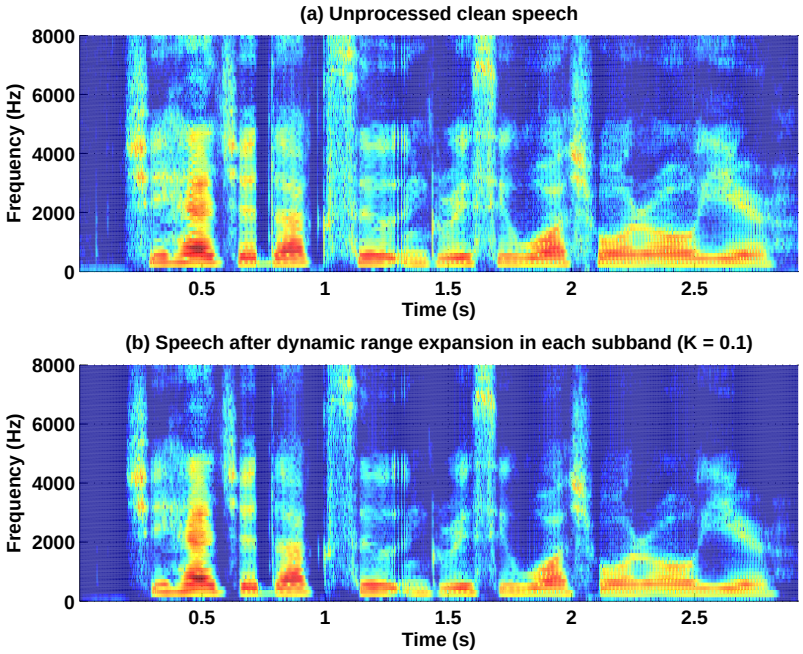


Figure 27: Spectrogram of clean speech before and after the dynamic range is expanded in each critical band $c[n]$.

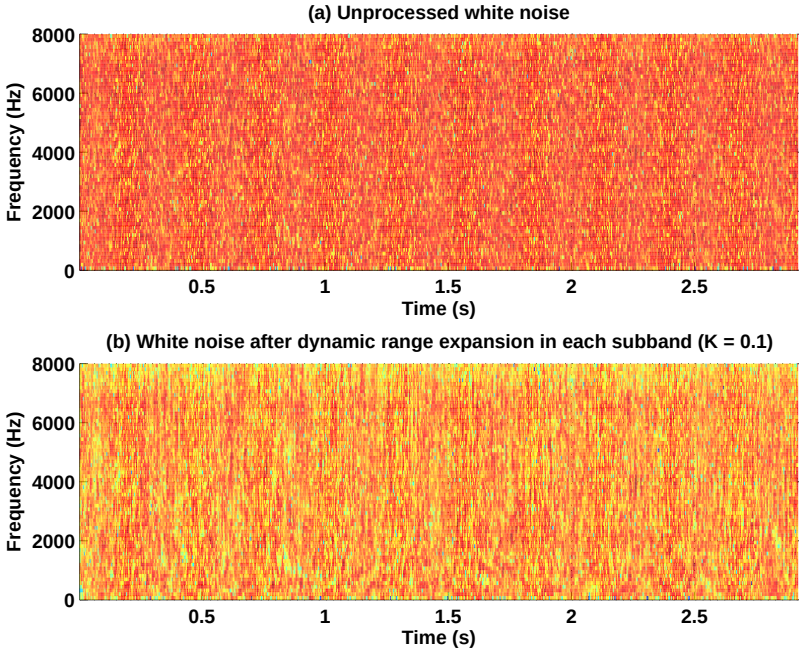


Figure 28: Spectrogram of white noise before and after the dynamic range is expanded in each critical band $c[n]$ using $e_{\mathrm{LPF}}[n]$

As described in Section 3.1, we are interested in processing the envelope that the human auditory systems maps to the loudness domain. However, there may not be a single correct

way to extract this envelope. In the next section, we investigate to see if the dynamic range of the signal is expanded in each critical band using the ideal envelope in the signal processing sense generates musical noise.

7.2 Dynamic Range Expansion of Noise using the Hilbert Envelope

According to the signal model presented in Section 3.2, each critical band is assumed to be a product of an envelope with an underlying carrier. Hence, each critical band can be assumed to be a modulated signal. To extract the envelope of a modulated signal exactly, either the carrier or the analytic signal is required. However, the envelope $e_{\mathrm{LPF}}[n]$ that we use in this research is only an approximate estimate of the envelope which may not necessarily be the true envelope of the signal in each critical band.

To investigate if estimate of the envelope $e_{\mathrm{LPF}}[n]$ is the cause of the musical noise, firstly, the ideal envelope is extracted. The envelope is ideal from a signal processing point of view and is obtained by computing the absolute value of the analytic signal of each critical band. Let us call this envelope $e_{\mathrm{H}}[n]$. The analytic signal of each critical band is formed using the Hilbert transform. The dynamic range of each critical band is expanded as described in Section 7.1. This dynamic range expansion using $e_{\mathrm{H}}[n]$ does not drastically change the perceptual quality of the noise. The spectrogram of white noise before and after the dynamic range expansion when $e_{\mathrm{H}}[n]$ is used in each critical band $c[n]$ is shown in Figure 29.

In the case where the dynamic range of the noise signal is expanded in each critical band using $e_{\mathrm{LPF}}[n]$, the musical noise is not heard in each individual critical band. The musical noise is heard when the signal is reconstructed by adding the processed critical bands to form the wide-band processed signal. This behavior leads us to believe that the mismatch in group delays of the IIR filters used to extract the critical bands may generate musical noise.

Figure 30 compares $e_{\mathrm{H}}[n]$ and $e_{\mathrm{LPF}}[n]$ for the critical band centered around 388 Hz. $e_{\mathrm{LPF}}[n]$ is extracted using a single-pole IIR filter. Hence, the envelope is not phase aligned with the corresponding critical band. There is no phase delay is between $e_{\mathrm{H}}[n]$ and the corresponding critical band $c[n]$. Moreover, the low cut-off frequency, in other words the

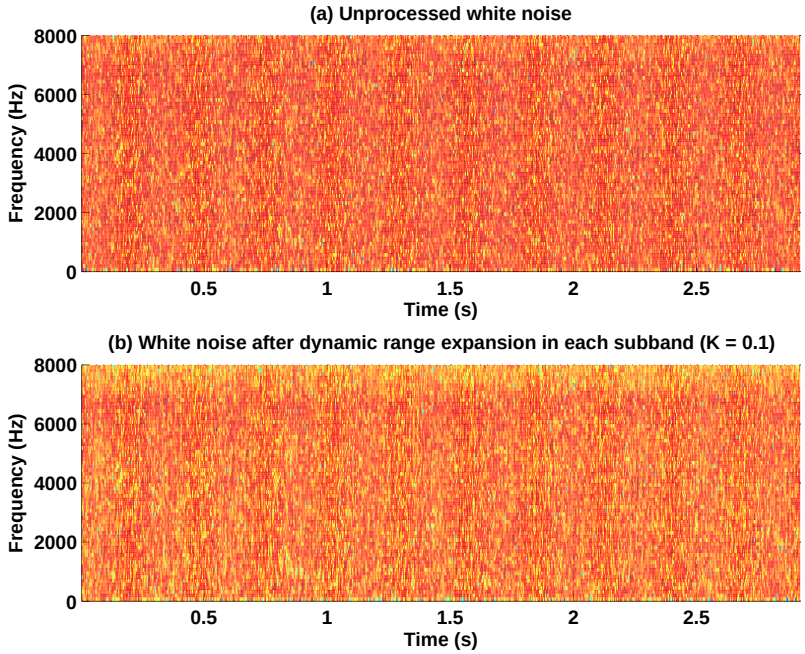


Figure 29: Spectrogram of white noise before and after the dynamic range expansion using $e_{\mathrm{H}}[n]$ in each critical band $c[n]$.

long time constant of the LPF restricts $e_{\mathrm{LPF}}[n]$ from rising or dropping to rapidly. However, since there is no constraint on the frequency content of $e_{\mathrm{H}}[n]$, the dynamic range of the $e_{\mathrm{H}}[n]$ is higher than $e_{\mathrm{LPF}}[n]$.

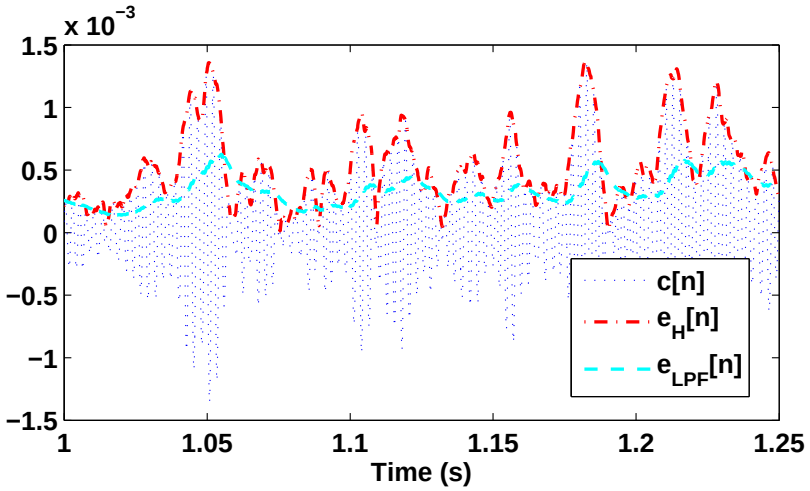


Figure 30: Critical band of white noise centered at 388 Hz and $e_{\mathrm{H}}[n]$ and $e_{\mathrm{LPF}}[n]$.

In the next sections, we will see how the

- Phase delay between critical bands

- Phase delay between the envelope and the corresponding critical band
- Dynamic range of the envelope

effect the dynamic range expansion and the amount of musical noise.

7.3 Phase Delay between Critical Bands

The IIR filters that are used to decompose the signal into critical bands do not have the same group delay across bands. Hence, the phase delay is not the same across critical bands. This phase delay mismatch combined with the dynamic range expansion may generate musical noise.

In [45], Petersen and Boll proposed a critical-bandwidth filter bank implementation, which ensures perfect phase and magnitude reconstruction. The Petersen-Boll filter bank decomposes the signal into critical bands that have a constant bandwidth for frequencies lower than 1 kHz, and a constant-Q factor for frequencies above 1 kHz. The analysis filter bank is a set of real-valued filters that have a zero response for the negative frequencies. Hence the output of the analysis filter bank is a complex time-varying signal, which is then complex demodulated to the baseband. The absolute value of this complex signal is the envelope $e_{\mathrm{PB}}[n]$ of the i-th critical band $c[n]$. The dynamic range of this envelope $e_{\mathrm{PB}}[n]$ is expanded in a similar way as described in Section 7.1 and then the signal is synthesized. The envelope expansion of this signal analysis and synthesis does not generate significant amount of musical noise.

If the inconsistent phase distortion across the critical is the cause of musical noise, then extracting the critical bands using a zero-phase or linear-phase distortion filter should also not generate musical noise. The `filtfilt` function of MATLAB filters a signal in both the forward and reverse direction ensuring the output has zero-phase distortion. The overall response of this filtering operation equals the squared magnitude of the original filter, as a result the effective cut-off frequency is lower than the original filter. An FIR filter has a linear-phase response. We can extract the critical bands using the `filtfilt` function or an FIR filter with the same cut-off frequencies as the IIR filters used in the filter bank. Instead of the single-pole IIR filter, we use `filtfilt` and FIR filters to extract the envelope to

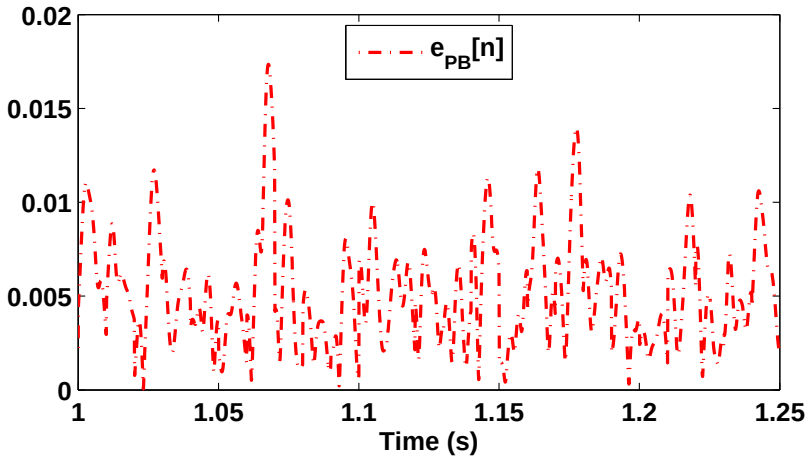


Figure 31: Envelope extracted using the Petersen-Boll critical band filter bank. The envelope is the absolute value of the complex critical band.

ensure the same phase distortion in each critical band. In each of case when the dynamic range of the envelope of the noise is expanded in each critical band, the processed wide-band output contains musical noise. Hence, the inconsistent phase distortion across the critical bands may not be the only cause of the musical noise.

7.4 Phase delay between the envelope and the corresponding critical band

The low pass filter that is used to extract the envelope $e_{\mathrm{LPF}}[n]$ is a single-pole IIR filter. Hence, there is a phase delay between $e_{\mathrm{LPF}}[n]$ and $c[n]$ for the i-th critical band. Such a phase delay is not present in $e_{\mathrm{H}}[n]$. This difference in phase between $e_{\mathrm{H}}[n]$ and $e_{\mathrm{LPF}}[n]$ can be seen in Figure 30. To remove this delay, we use the `filtfilt` function of MATLAB. Figure 32 shows the envelope $e_{\mathrm{ff}}[n]$ extracted using `filtfilt` function. The peaks of $e_{\mathrm{ff}}[n]$ are aligned with the peaks of the critical band, and is smoother compared to $e_{\mathrm{LPF}}[n]$. However, when we expand the dynamic range of each critical band of the noise using $e_{ff}[n]$, the output still contains musical noise.

The squaring effect of the magnitude response of the low-pass filter when using the `filtfilt` could be an added reason musical noise is still present when we use $e_{\mathrm{ff}}[n]$ to expand the dynamic range. We design a 128-tap linear-phase low-pass FIR filter with the same cut-off frequencies as mentioned in Section 3.2. The FIR-filter output was shifted

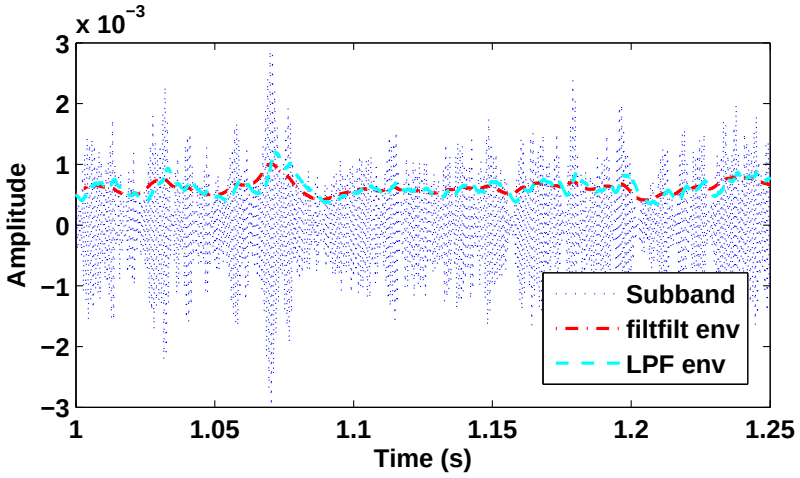


Figure 32: Subband of white noise centered at 1218 Hz and the corresponding envelopes extracted using `filtfilt` and LPF.

forward to compensate for the constant-phase delay. However, the output of the dynamic range expansion, in this case, also contains musical noise. Hence, the phase delay between the envelope and the critical band may not be the only cause of the musical noise.

7.5 Dynamic range of the envelope

A single-pole IIR filter is used to extract $e_{\mathrm{LPF}}[n]$. The cut-off frequency of these low pass filters are set to a fraction of the bandwidth of the critical band, as described in Section 3.2. Since the cut-off frequency of the filters are small (< 110 Hz), the filters have a long time constant. This long time constant limits how fast $e_{\mathrm{LPF}}[n]$ can follow the critical band, which restricts the dynamic range M of the envelope. Note, $M = \frac{e_{\max}}{e_{\min}} > 1$.

From equation (102), since $K < 1$, $\alpha > 1$, and as M decreases and approaches 1, α increases and approaches ∞. Let M_{LPF} be the dynamic range of the envelope $e_{\mathrm{LPF}}[n]$, and M_{H} of $e_{\mathrm{H}}[n]$. Since $M_{\mathrm{LPF}} < M_{\mathrm{H}}$, $\alpha_{\mathrm{LPF}} > \alpha_{\mathrm{H}}$. A higher value of α indicates a higher expansion ratio, which indicates a more aggressive dynamic range expansion. Figure 33 compares α_{LPF} and α_{H}.

The cut-off frequency of the low-pass filter could be increasaed to make sure that $e_{\mathrm{LPF}}[n]$ follows the critical band closely so that the dynamic range $M_{\mathrm{LPF}} \approx M_{\mathrm{H}}$, which would ensure that $\alpha_{\mathrm{LPF}} \approx \alpha_{\mathrm{H}}$. This increase in the frequency of the envelope reduces the musical noise when the dynamic range of the noisy-only signal is expanded, but deteriorates the quality

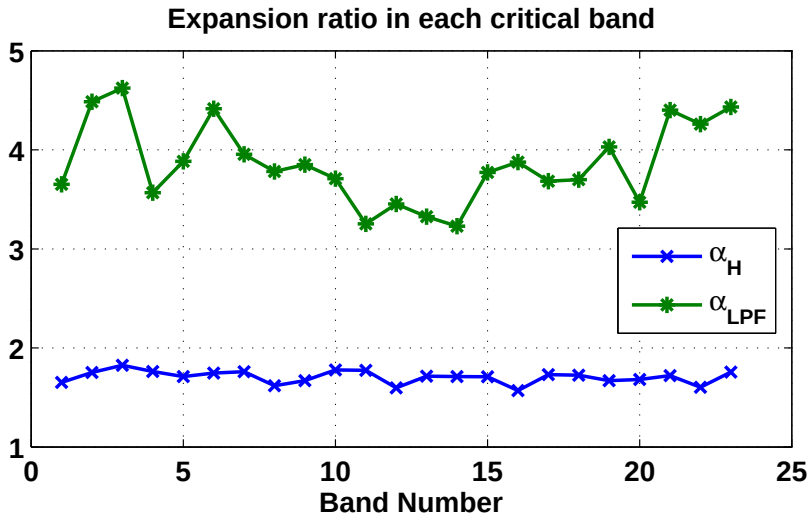


Figure 33: Comparing α calculated for each critical band when the dynamic range of white noise is expanded. The blue line shows the α_{H} calculated using $e_{\mathrm{H}}[n]$ and the green line shows the α_{LPF} calculated using $e_{\mathrm{LPF}}[n]$.

of speech when the dynamic range of speech is expanded. Hence, this approach to reduce the musical noise is not suitable.

As an experiment, to expand the dynamic range of the noisy-only signal in each critical band we use M_{H} to compute α_{H} in each critical band. In each critical band, β is computed as

$$\beta_{\mathrm{LPF}} = e_{\mathrm{max}}^{1-\alpha_{\mathrm{H}}}, \tag{107}$$

where $e_{\mathrm{max}} = \max(e_{\mathrm{LPF}}[n])$ in the i-th critical band. The dynamic range expansion gain $G[n]$ is computed as

$$G[n] = \beta_{\mathrm{LPF}}(e_{\mathrm{LPF}}[n])^{\alpha_{\mathrm{H}}-1}, \tag{108}$$

in each critical band. This gain $G[n]$ is applied to the corresponding critical band $c[n]$. In this case, the musical noise is drastically reduced, and remaining musical noise is mostly masked.

From Figure 33, we see that $1.5 < \alpha_{\mathrm{H}} < 2$. In the above experiment, using α_{H} to calculate the gain $G[n]$ based on $e_{\mathrm{LPF}}[n]$, the effective K is

$$K_{\mathrm{eff}} = \beta_{\mathrm{LPF}} e_{\mathrm{min}}^{\alpha_{\mathrm{H}}-1}, \tag{109}$$

where $e_{\mathrm{min}} = \min(e_{\mathrm{LPF}}[n])$. Since $\alpha_{\mathrm{H}} < \alpha_{\mathrm{LPF}}$, $K_{\mathrm{eff}} > K = 0.1$. Figure 35 shows K_{eff} for each

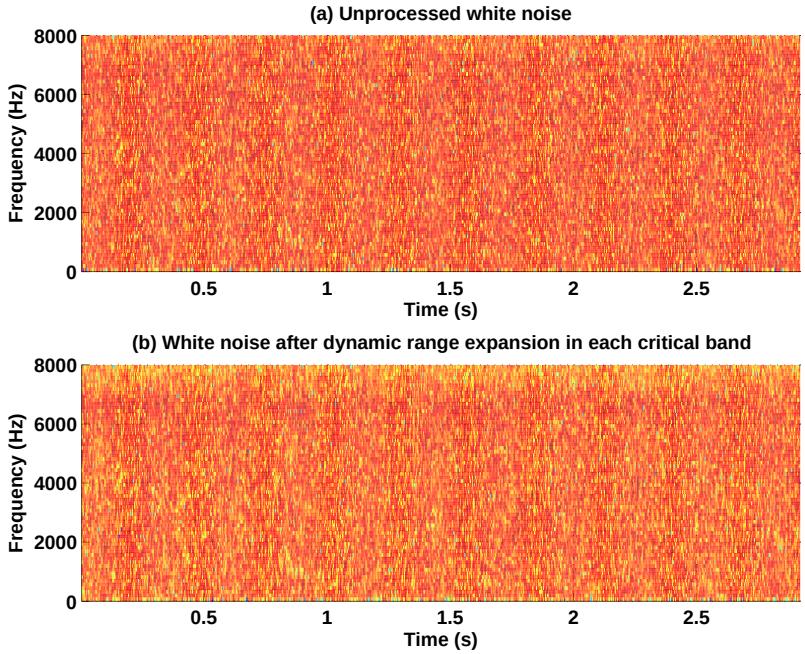


Figure 34: Spectrogram of white noise before and after the dynamic range is expanded in each critical band $c[n]$. The expansion ratio α is calculated using M_{H}, and β is calculated using $e_{\max} = \max(e_{\mathrm{LPF}}[n])$

critical band when α is calculated using M_{H}, and β is calculated using $e_{\max} = \max(e_{\mathrm{LPF}}[n])$. Hence, we can reduce the musical noise by limiting the expansion ratio of the envelope transformation.

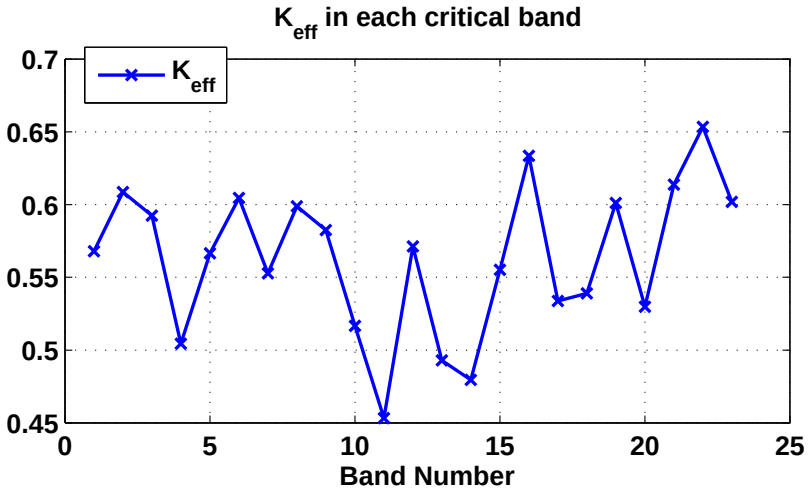


Figure 35: K_{eff} for each critical band when α is calculated using M_{H}, and β is calculated using $e_{\max} = \max(e_{\mathrm{LPF}}[n])$

7.6 Envelope Cut-off Frequency

As explained in Section 3.1, the envelope that is critical to the speech perception can be interpreted in different ways. Drullman *et al.* showed that the speech modulations in each critical band could be smoothed up to 16 Hz without reduction in speech intelligibility cite-Drullman1994. Ghitza showed that even if the speech modulations are smoothed to 16 Hz, the underlying carrier still contains information of the modulations [19]. Hence, if the noise-suppression gain is calculated from the an envelope that has a high cut-off frequency, it may combine with the underlying modulations present in the carrier to create audible artifacts.

We repeated the experiment described in Section 7.5, but this time we restrict the cut-off frequency of the envelope to 16 Hz in all critical bands to obtain the envelope $e_{16}[n]$. M_{H} is used to compute α_{H} in each critical band, β_{LPF} is calculated using to equation (107), and the dynamic range expansion gain $G[n]$ is calculated using to equation (108). The dynamic range expansion of the noise in each critical band in this case does not create any musical noise.

We repeat the same experiment for noisy speech signals. The cut-off frequency of the envelope is set to 16 Hz, M_{H} is used to compute α_{H}, and equations (107) and (108) to compute β_{LPF} and $G[n]$ respectively. The processed speech does not contain musical noise, and the quality of speech is intact. However, the noise suppression in the higher frequency bands is not significant. This degrade in the noise suppression performance is because β_{LPF} is computed such that the gain $G[n] = 1$ when $e[n] = e_{\max}$. In other words, the level $e_{\max}$ is not altered. In the high frequency critical bands where the level of the speech information is low, the critical band may be dominated by the noise and our assumption that $e_{\max}$ corresponds to the speech level may not be true. In this case, the noise suppression gain is such that it preserves the noise level, and hence the noise-suppression performance degrades. Figure 36

7.7 Expansion Ratio

As seen in Section 7.5 and 7.6, the expansion ratio of envelope transform and the cut-off frequency of the envelope determine the amount of musical noise that is perceived when the

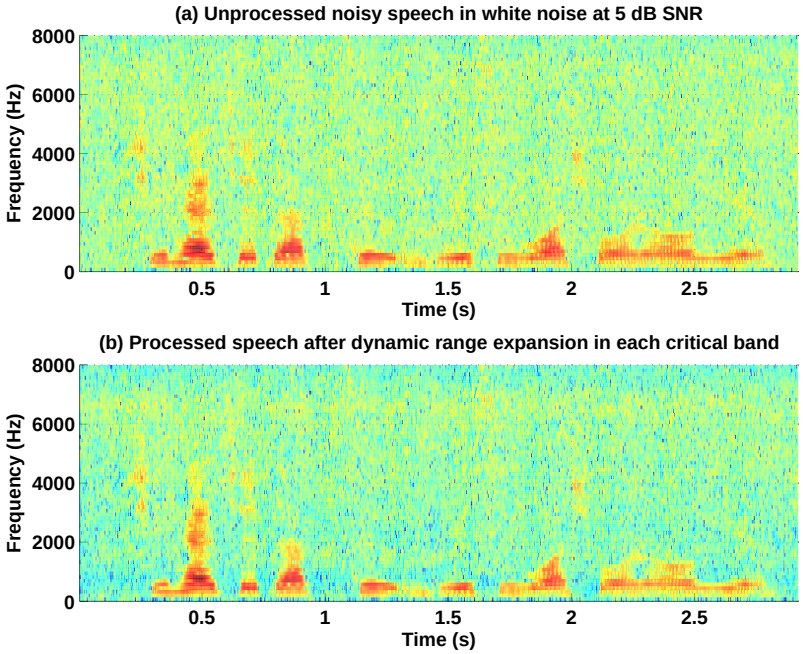


Figure 36: Spectrogram of speech in white noise at 5 dB SNR before and after the dynamic range is expanded in each critical band $c[n]$. The cut-off frequency of the envelope is limited to 16 Hz. The expansion ratio α is calculated using M_{H}, and β is calculated using $e_{\max} = \max(e_{16}[n])$

dynamic range of the envelope is expanded to suppress the background noise. Firstly, in this section, we find the maximum expansion ratio that does not create musical noise when the dynamic range of the envelope is expanded.

To design an experiment to determine the maximum expansion ratio $\alpha_{\max}$ in each critical band, we turn to experiments that have been performed for digital hearing aids. One of the causes of hearing loss is the due to the loss in the outer hair cells' ability to non-linearly compress the dynamic range of the input signal. In digital hearing aids, the dynamic range of the signal can be compressed to compensate for the hearing loss. The dynamic range is logarithmically compressed using a similar form as equation (13). However, to obtain compression K is set to $K_{\mathrm{comp}} > 1$, and the parameters β and α are calculated accordingly. Studies have been done to study the effect of multi-band dynamic range expansion on the speech intelligibility and speech quality [28, 47, 55]. Van Buuren *et al.* evaluated the speech intelligibility and sound quality for different compression ratios and expansion ratios, processing the signal in different number of bands [55]. However, the compression

and expansion ratios that were tested were greater than 2.

The dynamic range compression in hearing aids, and the dynamic range expansion in our noise-suppression system mimics the non-linear processing of the outer hair cells. We expect the compression/expansion ratios that would not degrade the quality of the processed speech depends on the compression ratio of the outer hair cells. Keeping these facts in mind, we can design an experiment to find the maximum expansion ratio that does not generate musical noise when the dynamic range of noisy speech is expanded to suppress noise.

7.7.1 Experimental Setup

To find the maximum expansion ratio that does not create audible artifacts in the processed signal, we process the noise-only signal. If we processed noisy speech signal to find the expansion ratio threshold, the speech may mask some of the audible artifacts, which may result in an erroneous threshold.

Most hearing aids operate in 3 frequency groups—<1 kHz, $1-2.5$ kHz, and >2.5 kHz. To avoid adjusting the expansion ratio in each of the 23 critical bands that are obtained at 16 kHz sampling rate, we vary the expansion ratio α by the same amount in each of the critical bands that lie within the frequency groups. For a frequency group, the critical bands are extracted using 4-th order IIR filters, which have been described in Section 3.2. Each critical band in the frequency group is full-wave rectified and low-pass filtered to extract the envelope. The cut-off frequency of the low-pass filter to extract the envelope in each critical band is set to 16 Hz.

The gain $G[n]$ that expands the dynamic range of the signal in each critical band is calculated using equation (105). α is varied between $1-2$ in steps of 0.1, and β is calculated using equation (16). The gain $G[n]$ is applied to each critical band, and the critical bands present in the frequency group being tested are summed to form the processed signal. The value of α is increased till musical noise was heard in the processed output. The maximum expansion ratio α that does not generate musical noise is the processed noise signal $\alpha_{\max}$ for the different frequency groups tested is listed in Table 12.

Table 12: Maximum expansion ratio $\alpha_{\max}$ for the critical bands in each frequency group.

Frequency Group (kHz)	$\alpha_{\max}$
< 1	1.3
$1 - 2.5$	1.9
> 2.5	1.9

7.8 Improved noise-suppression algorithm

Taking into account what we learned about musical noise from this chapter, we can improve the performance of the perceptual noise suppression system described in Chapter 3. We briefly describe this system here.

The noisy signal $x[n]$ is split into critical bands $c[n]$ using the filter bank described in Section 3.2. Note, the subscript i that indicates the critical band number is dropped for convenience. In each critical band, the envelope $e[n]$ is extracted by full-wave rectifying and low-pass filtering the critical band $c[n]$. The cut-off frequency of the envelope $e[n]$ is set to 16 Hz. The maximum $e_{\max}$ and minimum $e_{\min}$ of the envelope in each critical band is computed.

The expansion ratio α is computed using equation (102), where $K = 0.1$ and $M = \frac{e_{\max}}{e_{\min}}$. The expansion ratio is then modified based on the following conditions.

$$\alpha = \begin{cases} \alpha, & \text{if } \alpha < \alpha_{\max} \\ \alpha_{\max}, & \text{otherwise.} \end{cases} \tag{110}$$

Equation (16) is used to calculate the value for β. As explained in Section 7.6, β is calculated such that the signal level $e_{\max}$ is not altered. At higher frequencies (> 3.5 kHz), the noise signal is typically dominant over the speech signal, hence $e_{\max}$ may not correspond to the signal level. To increase the performance of noise suppression, for the critical bands whose center frequency is > 3.5 kHz, we can calculate β using $e_{\min}$ as

$$\beta = \begin{cases} e_{\max}^{1-\alpha}, & \text{if } f_{\text{CB}} < 4\text{kHz} \\ \gamma e_{\min}^{1-\alpha}, & \text{otherwise,} \end{cases} \tag{111}$$

where $\gamma < 1$ and is set to 0.2. γ determines how much the noise is suppressed in the high

frequency bands.

Figures 37 and 38 compares the spectrograms of speech present in white noise and pub noise at 5 dB SNR before and after noise suppression respectively.

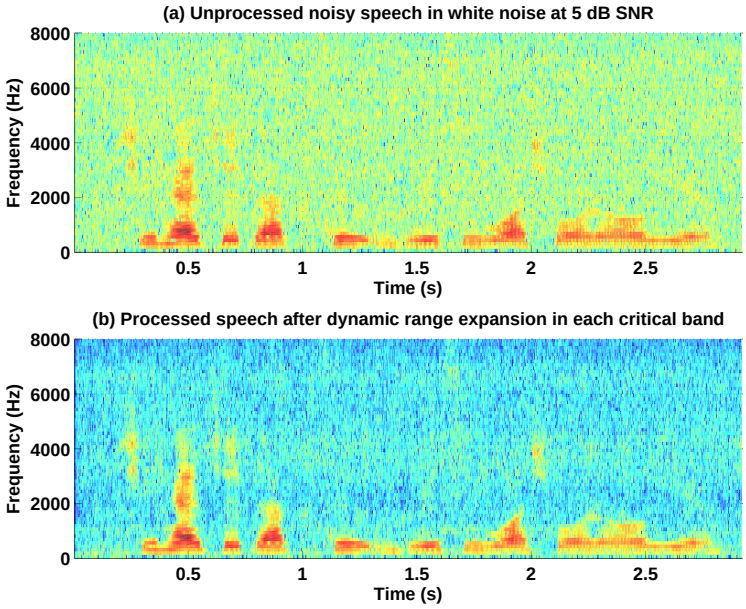


Figure 37: Spectrogram of speech in white noise at 5 dB SNR before and after the improved perceptual noise suppression.

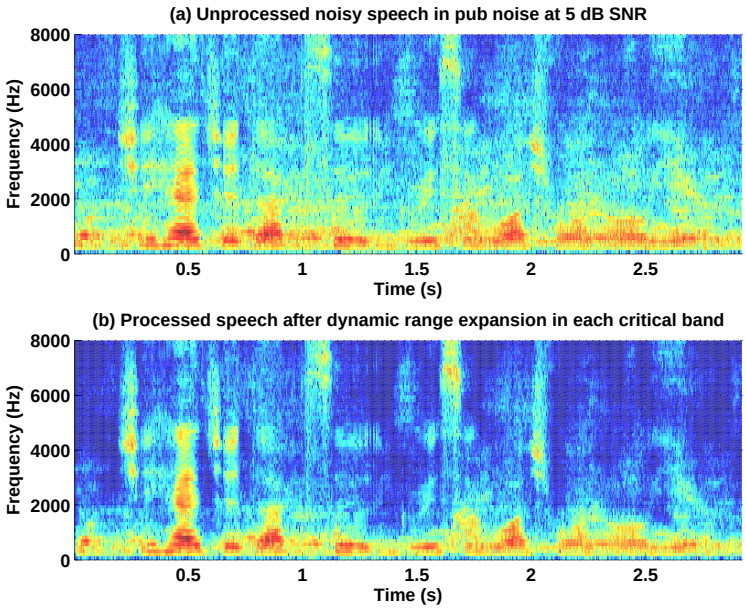


Figure 38: Spectrogram of speech in pub noise at 5 dB SNR before and after the improved perceptual noise suppression.

CHAPTER VIII

CONCLUDING REMARKS

In this thesis, we step away from the traditional methods of noise suppression that are based on mathematical optimal estimators and develop a noise-suppression paradigm that is based on a model of the human auditory system. In this approach, we can process speech signals in the perceptual domain, and hence process signals in way that is natural to the human auditory perceptual system. When the processed signal is meant to be heard by the human ear, noise-suppression algorithms based on this perceptual paradigm allows us to suppress the background noise present in speech without altering the perceived quality of speech. Moreover, we can control the processing so that the audible artifacts are reduced. Through subjective tests, we show that this approach out-performs traditional noise-suppression techniques in terms of the quality of speech.

8.1 List of Contributions

We summarize the major contributions of this thesis below:

Single-channel noise-suppression system We develop a noise-suppression system in which the signal to be processed is analysed based on the frequency decomposition of the cochlea. Moreover, the noise-suppression gain is based on the non-linear processing of the perceptual auditory system. The output enhanced speech of such processing sounds natural when heard by the human ear.

Multi-microphone noise-suppression system We use well-known source-separation methods in conjunction with the perceptual noise-suppression system to further improve the noise-suppression performance. The source-separation algorithm provides an estimate of the noise present in the noisy speech, which is then used to suppress the noise. Hence, the perceptual noise-suppression system serves as a source-separation post-processing system.

Real-time implementation of noise-suppression system Moving towards a realistic implementation of the perceptual post processing for source separation, we come across issues that result in audible artifacts in the processed output in the form of musical noise. We address this issue, and hence reduce the perception of the musical noise by implicitly smoothing the noise-suppression gain over time.

Optimal-estimation techniques in the perceptual domain Using optimal-estimation techniques in the perceptual domain, we show that us to estimate the noise-suppression gain parameters systematically.

Understanding the modulation artifacts By understanding which parameters are crucial to the perception of artifacts, we can constrain the noise-suppression gain to reduce the audible artifacts in the processed signal.

8.2 Future Work

Signal Models The carrier may be modelled as an FM signal rather than a single frequency signal. This model may lead to a better understanding of how to extract the envelope that can be then processed without generating any audible artifacts.

Model product terms Mathematically modelling the product terms that arise from processing the critical bands, may lead to a better understanding of what type of processing would or would not create audible artifacts. This understanding will allow us to develop better speech and audio processing algorithms.

Source separation The human auditory systems outperforms existing techniques of source localization and separation. By understanding how the auditory system separates and focus on a particular source when it present in other competing sources, we can develop source separation systems that may outperform existing techniques.

REFERENCES

[1] Alam, M., O'Shaughnessy, D., and Selouani, S.-A., "A new perceptual post-filter for single channel speech enhancement," in *Electrical and Computer Engineering (ICECE), 2008 International Conference on*, pp. 386 –390, Dec. 2008.

[2] Amari, S., Douglas, S., Cichocki, A., and Yang, H., "Multichannel blind deconvolution and equalization using the natural gradient," in *Signal Processing Advances in Wireless Communications, 1997 IEEE Signal Processing Workshop on*, pp. 101–104, IEEE, 1997.

[3] Anderson, D., "Model based development of a hearing aid," Master's thesis, Brigham Young University, Provo, Utah, Apr. 1994.

[4] Anderson, D., *Audio signal noise reduction using multi-resolution sinusoidal modeling*. PhD thesis, Georgia Institute of Technology, Atlanta, Georgia, Mar. 1999. Chapter 5.

[5] Anderson, D., "A modulation view of audio processing for reducing audible artifacts," in *Acoustics, Speech, and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 5474 –5477, Mar. 2010.

[6] Araki, S., Makino, S., Hinamoto, Y., Mukai, R., Nishikawa, T., and Saruwatari, H., "Equivalence between Frequency-Domain Blind Source Separation and Frequency-Domain Adaptive Beamforming for Convolutive Mixtures," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 11, pp. 1157–1166, 2003.

[7] Araki, S., Makino, S., Sawada, H., and Mukai, R., "Reducing Musical Noise by a Fine-shift Overlap-add Method Applied to Source Separation Using a Time-frequency Mask," in *Acoustics, Speech, and Signal Processing (ICASSP), 2005 IEEE International Conference on*, pp. 81–84, IEEE, 2005.

[8] Bell, A. and Sejnowski, T. J., "An information-maximization approach to blind separation and blind deconvolution.," *Neural computation*, vol. 7, pp. 1129–59, Nov. 1995.

[9] Berouti, M., Schwartz, R., and Makhoul, J., "Enhancement of speech corrupted by acoustic noise," in *Acoustics, Speech, and Signal Processing (ICASSP), 1979 IEEE International Conference on*, vol. 4, pp. 208 – 211, Apr. 1979.

[10] Boll, S., "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, pp. 113 – 120, Apr. 1979.

[11] Capdevielle, V., Serviere, C., and Lacoume, J., "Blind separation of wide-band sources in the frequency domain," in *Acoustics, Speech, and Signal Processing (ICASSP), 1995 IEEE International Conference on*, pp. 2080–2083, IEEE, 1995.

[12] Cappe, O., "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, pp. 345 –349, apr 1994.

[13] Chabries, D., Anderson, D., Stockham, T.G., J., and Christiansen, R., "Application of a human auditory model to loudness perception and hearing compensation," in *Acoustics, Speech, and Signal Processing (ICASSP), 1995 International Conference on*, vol. 5, pp. 3527 –3530 vol.5, May 1995.

[14] Christiansen, M. W., *Digital speech processing in the context of a human auditory model*. PhD thesis, Brigham Young University, Provo, Utah, 1990.

[15] Comon, P., "Independent component analysis, A new concept?," *Signal Processing*, vol. 36, pp. 287–314, Apr. 1994.

[16] Douglas, S., Cichocki, A., and Amari, S., "Multichannel blind separation and deconvolution of sources with arbitrary distributions," in *Neural Networks for Signal Processing, 1997 IEEE Signal Processing Society Workshop on*, pp. 436–445, IEEE, 1997.

[17] Drullman, R., Festen, J. M., and Plomp, R., "Effect of temporal envelope smearing on speech reception.," *The Journal of the Acoustical Society of America*, vol. 95, pp. 1053–64, Feb. 1994.

[18] Ephraim, Y. and Malah, D., "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, pp. 1109 – 1121, Dec. 1984.

[19] Ghitza, O., "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception," *The Journal of the Acoustical Society of America*, vol. 110, p. 1628, 2001.

[20] Gustafsson, S., Jax, P., Kamphausen, A., and Vary, P., "A postfilter for echo and noise reduction avoiding the problem of musical tones," in *Acoustics, Speech, and Signal Processing (ICASSP) 1999 IEEE International Conference on*, vol. 2, pp. 873 –876 vol.2, Mar. 1999.

[21] Gustafsson, S., Jax, P., and Vary, P., "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics," in *Acoustics, Speech and Signal Processing (ICASSP), 1998 IEEE International Conference on*, May 1998.

[22] Hansen, J., "Morphological constrained feature enhancement with adaptive cepstral compensation (mce-acc) for speech recognition in noise and lombard effect," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, pp. 598 –614, oct 1994.

[23] Haykin, S., *Unsupervised Adaptive Filtering, Volume 1: Blind Source Separation*. John Wiley & Sons, Inc., 2000. Chapter 8.

[24] Houtgast, T. and Steeneken, H., "A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria," *The Journal of the Acoustical Society of America*, vol. 77, no. 3, pp. 1069–1077, 1985.

[25] Ikram, M. and Morgan, D. R., "A Beamforming approach to permutation alignment for multi-channel frequency domain blind source separation," *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, no. 2, pp. 881–884, 2002.

[26] Johnston, J., "Transform coding of audio signals using perceptual noise criteria," *Selected Areas in Communications, IEEE Journal on*, vol. 6, pp. 314 –323, Feb. 1988.

[27] Kamath, S. and Loizou, P. C., "A postfilter for echo and noise reduction avoiding the problem of musical tones," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 4, May 2002.

[28] Kates, J., *Digital hearing aids*. Cambridge Univ Press, 2008.

[29] Kolossa, D. and Orglmeister, R., "Nonlinear postprocessing for blind speech separation," in *Proc. of Independent Component Analysis and Blind Signal Separation (ICA '04)*, 2004.

[30] Lai, Y.-P., Hui, M.-C., Kok, C.-W., and Siu, M.-H., "Speech recognition enhancement by psychoacoustic modeled noise suppression," in *Multimedia and Expo (ICME), 2004 IEEE International Conference on*, vol. 2, pp. 1335 –1338 Vol.2, June 2004.

[31] Li, Y., Cichocki, A., and Zhang, L., "Blind source estimation of FIR channels for binary sources: a grouping decision approach," *Signal Processing*, vol. 84, pp. 2245–2263, Dec. 2004.

[32] Lim, J. and Oppenheim, A., "All-pole modeling of degraded speech," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, pp. 197 – 210, June 1978.

[33] Loizou, P., *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[34] Makino, S., Sawada, H., Mukai, R., and Araki, S., "Blind source separation of convolutive mixtures of speech in frequency domain," *Fundamentals of Electronics, Communications and Computer Sciences, IEICE Transactions on*, vol. 88, no. 7, p. 1640, 2005.

[35] Marin-Hurtado, J. and Anderson, D., "Fft-based block processing in speech enhancement: Potential artifacts and solutions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, pp. 2527 –2537, nov. 2011.

[36] Moore, B., *An introduction to the psychology of hearing*. Emerald Group Pub Ltd, 2003.

[37] Noohi, T. and Kahaei, M., "Residual cross-talk suppression for convolutive blind source separation," in *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*, vol. 1, pp. V1–543 –V1–547, Apr. 2010.

[38] Papoulis, A. and Pillai, S. U., *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 2000.

[39] Parikh, D. and Anderson, D., "Blind source separation with perceptual post processing," in *IEEE Digital Signal Processing and Signal Processing Education Workshop*, 2011.

[40] Parikh, D., Ravindran, S., and Anderson, D., "Gain adaptation based on signal-to-noise ratio for noise suppression," in *Applications of Signal Processing to Audio and Acoustics (WASPAA). 2009 IEEE Workshop on*, pp. 185 –188, Oct. 2009.

[41] Park, K. S., Park, J., Son, K., and Kim, H. T., "Postprocessing with wiener filtering technique for reducing residual crosstalk in blind source separation," *Signal Processing Letters, IEEE*, vol. 13, pp. 749 –751, Dec. 2006.

[42] Parra, L. and Spence, C., "Convolutive blind separation of non-stationary sources," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, pp. 320 –327, May 2000.

[43] Pedersen, M. S., Larsen, J., Kjems, U., Parra, L. C., and Lyngby, K., "A survey of convolutive blind source separation methods," *Speech Communication*, pp. 1–34, 2007.

[44] Petersen, T. and Boll, S., "Acoustic noise suppression in the context of a perceptual model," in *Acoustics, Speech, and Signal Processing (ICASSP), 1981 IEEE International Conference on*, vol. 6, pp. 1086–1088, IEEE, 1981.

[45] Petersen, T. and Boll, S., "Critical band analysis-synthesis," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 31, pp. 656 – 663, jun 1983.

[46] Plomp, R., "Perception of speech as a modulated signal," in *Proceedings of the Tenth International Congress of Phonetic Sciences*, pp. 29–40, Dordrecht, Foris, 1983.

[47] Plomp, R., "The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function," *The Journal of the Acoustical Society of America*, vol. 83, p. 2322, 1988.

[48] Ravindran, S., *Physiologically Motivated Methods For Audio Pattern Classification*. PhD thesis, Georgia Institute of Technology, Nov. 2006.

[49] Rosca, J., Balan, R., and Beaugeant, C., "Multi-channel psychoacoustically motivated speech enhancement," in *Acoustics, Speech, and Signal Processing (ICASSP), 2003 IEEE International Conference on*, vol. 1, pp. I – 84–7 vol.1, Apr. 2003.

[50] Saruwatari, H., Kurita, S., Takeda, K., Itakura, F., Nishikawa, T., and Shikano, K., "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 1, p. 1146, 2003.

[51] Scalart, P. and Filho, J., "Speech enhancement based on a priori signal to noise estimation," in *Acoustics, Speech, and Signal Processing (ICASSP), 1996 IEEE International Conference on*, vol. 2, pp. 629 –632 vol. 2, May 1996.

[52] Takahashi, Y., Takatani, T., Osako, K., Saruwatari, H., and Shikano, K., "Blind spatial subtraction array for speech enhancement in noisy environment," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, pp. 650 –664, May 2009.

[53] Thiemann, J. and Kabal, P., "Low distortion acoustic noise suppression using a perceptual model for speech signals," in *Speech Coding, 2002, IEEE Workshop Proceedings.*, Oct. 2002.

[54] Torkkola, K., "Blind separation for audio signals-are we there yet?," in *Proc. of Independent Component Analysis and Blind Signal Separation (ICA '01)*, vol. 1, Citeseer, 1999.

[55] van Buuren, R., Festen, J., and Houtgast, T., "Compression and expansion of the temporal envelope: Evaluation of speech intelligibility and sound quality," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2903–2913, 1999.

[56] Viemeister, N., "Temporal modulation transfer functions based upon modulation thresholds," *The Journal of the Acoustical Society of America*, vol. 66, no. 5, pp. 1364–1380, 1979.

[57] Virag, N., "Single channel speech enhancement based on masking properties of the human auditory system," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, pp. 126 –137, Mar. 1999.

[58] Yost, W. A., *Fundamentals of Hearing: an Introduction*. Academic Press, 1997.

[59] Zwicker, E. and Fastl, H., *Psychoacoustics*. New York: Springer-Verlag, 1990.

[60] Zwillinger, D. and Jeffrey, A., *Table of Integrals, Series, and Products*. Academic Press, 2007. Section 3.462.

VITA

Devangi Nikunj Parikh was born on May 29, 1984 in Joplin, Missouri and graduated from St. Xavier's Loyola Hall High School in 2002. She received her B.E. in Electronics and Communication Engineering from Nirma Institute of Technology in 2006. She joined Georgia Institute of Technology in 2006 as a graduate student. Currently, her research interests are in the general area of speech and audio processing. She is interested in understanding the human perceptual auditory model and applying this model to noise suppression algorithms to obtain perceptually pleasing speech quality. Her research interests also include single- and multi-channel speech enhancement and noise suppression.